2016

# On the molecular biology and evolution of plant parasitism by nematodes

Jason Brett Noon
*Iowa State University*

# On the molecular biology and evolution of plant parasitism by nematodes

by

**Jason Brett Noon**

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics and Genomics

Program of Study Committee:
Thomas Baum, Major Professor
Yanhai Yin
Drena Dobbs
Bing Yang
Erik Vollbrecht

Iowa State University

Ames, Iowa

2016

# TABLE OF CONTENTS

## ACKNOWLEDGMENTS

I would like to thank my committee chair, Thomas Baum, and my committee members, Yanhai Yin, Drena Dobbs, Bing Yang, and Erik Vollbrecht, for their guidance and support throughout the course of this research.

In addition, I would also like to thank my friends, colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience. I want to also offer my appreciation to those who were willing to assist in the research included in this dissertation. Most of all, I would like to thank my wife Iris, without whom, this dissertation would not have been possible.

## ABSTRACT

Plant-parasitic nematodes (PPN) are among the most devastating plant pathogens. However, our understanding of how nematodes adapted to plant parasitism, and the molecular mechanisms that PPN use during infection is limited. Among the most important genomic changes that occurred in the free-living nematode ancestors of PPN were multiple horizontal gene transfer (HGT) events from bacteria. Though it is clear that HGT helped shape the genomes of many PPN, how this process occurred is unknown. Also, it is evident that successful parasitism occurs from the delivery of proteinaceous effectors into plant roots to hijack and modify host cellular processes.

The research included in this dissertation aims at addressing several important questions regarding HGT in PPN, and investigates important molecular, cellular and developmental processes that are determined critical for successful parasitism. Of particular emphasis throughout this dissertation is the soybean cyst nematode, *Heterodera glycines*, due to a highly specialized and agronomically important interaction with its soybean host.

Major findings for HGT in PPN include the identification of eighteen new *H. glycines* effectors, three of which are determined to have been part of more ancient HGT events from rhizosphere bacteria. Additionally, homologs of two of the three HGT genes are shown to have been transferred numerous different times from bacteria to diverse eukaryotes and archaea. The latter findings indicate the likely evolutionary advantages that these genes provided not just to PPN, but many different taxa. Intriguingly, we reveal that a group of retroviruses specific to distal nematode clades is genomically associated

with HGT genes in PPN genomes. These retroviruses potentially have all of the elements that would be necessary for HGT to occur in PPN. Thus, we propose the tempting hypothesis that this specific group of retroviruses might have contributed to HGT in these nematodes.

We also reveal several novelties for plant-nematode interactions. Major findings include the discovery of a strongly expressed *H. glycines* effector that is essential for virulence and efficiently targets plant cell nucleoli for suppression of innate immune responses. Also, this *H. glycines* effector contains marginal, but significant sequence similarity with an immunosuppressive effector found only in *Plasmodium* spp., the malaria parasites. Extensive database searches, phylogenetic analyses, and functional complementation experiments conclude that the similarities are best explained by sequence convergence due to similar immunosuppressive functions. Furthermore, we determine that a specific microRNA network in soybean that is essential for plant development delineates the formation of the *H. glycines* feeding site, and interfering with this network renders soybean roots much less susceptible to infection.

In conclusion, the major findings included in this dissertation reveal novel insights into how nematodes adapted to plant parasitism, and for how PPN manipulate their host plants during infection to establish compatible interactions. Moreover, these findings will undoubtedly provide foundations for developing novel control measures against these important plant pathogens.

# CHAPTER 1. GENERAL INTRODUCTION

## 1.1 Nematode adaptation to plant parasitism

The phylum Nematoda is incredibly speciose with 27,000 documented species, and nematodes (roundworms) are ubiquitous in soils, and in freshwater and marine sediments (Quist et al. 2015). This phylum has most recently been divided into 12 different clades (Fig. 1) based on phylogenetic data from small subunit ribosomal DNA (SSU rDNA) sequences (Holterman et al. 2006). Both morphological and phylogenetic data strongly suggest that plant parasitism evolved at least four independent times in Nematoda. An estimated 440 million years ago (mya), the first diversification event in Nematoda took place resulting in the split between the subclasses Enoplia (clade 1), and Dorylaimia and Chromadoria (Fig. 1). Many millions of years later, Dorylaimia (clade 2) and Chromadoria (clades 3-12) split, thus leading to the three independent subclasses (Fig. 1). Then, long after the diversification of many independent lineages, plant parasites emerged within the Trichodoridae and Longidoridae families of Enoplia and Dorylaimia, respectively (Fig. 1). Interestingly, in the distal Chromadoria, at least two independent cases of plant parasitism emerged (Fig. 1). The first occurred within the clade 10 families Aphelenchoididae and Parasitaphelenchidae (possibly their unique common ancestor). The second occurred within the clade 12 order Tylenchida. The Hoplolaimina suborder within Tylenchida (Fig. 1) contains the plant parasites of highest agronomic importance (Quist et al. 2015).

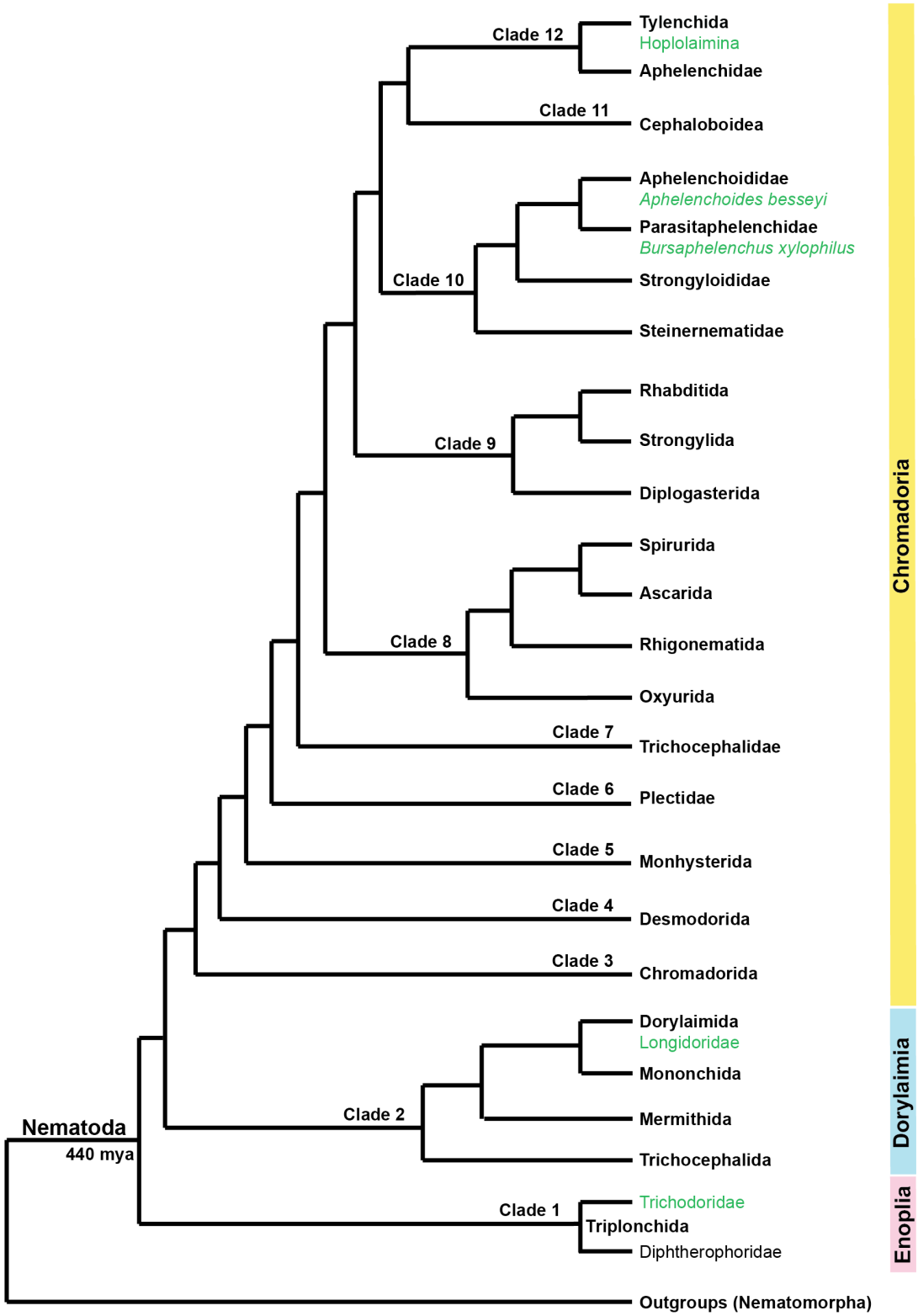

Tylenchida
Hoplolaimina
Aphelenchidae
Clade 12
Clade 11
Cephaloboidea
Aphelenchoididae
Aphelenchoides besseyi
Parasitaphelenchidae
Bursaphelenchus xylophilus
Strongyloididae
Clade 10
Steinernematidae
Rhabditida
Strongylida
Clade 9
Diplogasterida
Spirurida
Ascarida
Rhigonematida
Clade 8
Oxyurida
Clade 7
Trichocephalidae
Clade 6
Plectidae
Clade 5
Monhysterida
Clade 4
Desmodorida
Clade 3
Chromadorida
Dorylaimida
Longidoridae
Mononchida
Clade 2
Mermithida
Trichocephalida
Nematoda
440 mya
Trichodoridae
Clade 1
Triplonchida
Diphtherophoridae
Outgroups (Nematomorpha)
Chromadoria
Dorylaimia
Enoplia

**Fig. 1. Cladogram of the phylum Nematoda.** The three Nematoda subclasses are illustrated to the right of the corresponding clades. Plant-parasitic nematode lineages are illustrated in green. This figure is adapted from (Haegeman et al. 2011; Quist et al. 2015).

Interestingly, phylogenetic data from SSU rDNA sequences strongly suggest that each of the four major lineages of plant parasites evolved from fungivores since they cluster together in monophyletic groups. These conclusions are supported by morphological data, in particular the presence and structure of cuticularized puncturing devices [style(t)s] in the head region for penetrating cell walls. For instance, the Trichodoridae plant parasites from clade 1 (Fig. 1) have a curved protrusible onchiostyle that consists of a grooved solid tooth (onchium) connected to an onchiophore. Accordingly, the fungivores within the suspected sister lineage of Trichodoridae (Diphtherophoridae; Fig. 1) also have an onchiostyle. Also, the Longidoridae plant parasites from clade 2 (Fig. 1) have a spear-like odontostyle connected to an odontophore, and their fungivorous relatives within the same order, Dorylaimida, also have odontostyles. Furthermore, both clade 10 Aphelenchoididae/Parasitaphelenchidae and clade 12 Hoplolaimina plant parasites, as well as their sister lineages of fungivores, are equipped with stomatostylets (or simply stylets), though they are smaller in fungivores. The stylet consists of a hollow conus and shaft typically with three knobs at the base (Quist et al. 2015).

Another line of evidence that supports plant parasitism having evolved from fungivory is the presence and conservation of specific families of cell-wall modifying proteins (CWMPs) known as cellulases (β-1,4-endoglucanases). Cellulases are conserved within some clades of plant parasites and their closely related fungivores. For example, Hoplolaimina plant parasites and their closely related fungivores within clade 12 (Fig. 1,

family Aphelenchidae) contain glycoside hydrolase family 5 (GH5) cellulases. Also, within Parasitaphelenchidae and Aphelenchoididae, *Bursaphelenchus xylophilus* and *Aphelenchoides besseyi*, respectively (Fig. 1), as well as the congeneric fungivorous relatives of these two species, have GH45 cellulases. However, although the Longidoridae plant parasites (Fig. 1) have cellulases from another glycoside hydrolase family, these CWMPs have not been documented in their closely releated fungivores. Moreover, cellulases have not been documented in the Trichodoridae plant parasites (Quist et al. 2015). Thus, although cellulases support the evolution of plant parasitism from fungivory, in particular in Chromadoria clades 10 and 12, the appearance of these CWMPs in the different subclasses may have occurred through different courses.

Several independent lines of evidence indicate that multiple genes (including the *GH5 cellulases* mentioned above) in plant-parasitic nematodes, in particular the Hoplolaimina plant parasites (Fig. 2), are most likely derived from bacteria. This is because all CWMPs in Hoplolaimina plant parasites such as GH5 cellulases, xylanases, arabinases, pectate lyases, polyglacturonases and expansins are most similar to sequences from plant-pathogenic, symbiotic or saprophytic soil bacteria. Also, in phylogenetic analyses, the CWMP sequences from these nematodes and bacteria form well-supported clusters, and statistical methods rule out decent via common ancestry in eukaryotes (Danchin et al. 2010; Haegeman et al. 2011). The same can be stated for the presence of almost an entire vitamin B biosynthetic and salvage pathway in cyst nematodes (family Heteroderidae; Fig. 2) (Craig et al. 2008; Craig et al. 2009), and other genes [reviewed in (Haegeman et al. 2011)]. Currently, the most likely scenario to explain these findings is that HGT of CWMPs and possibly other genes occurred in bacteriovores that led to the

emergence of the ancestral fungivores, and was followed by the emergence of the Hoplolaimina plant parasites (Quist et al. 2015). This scenario is supported by the position of the bacteriovorous superfamily Cephaloboidea basal to Tylenchida and the fungivorous Aphelenchidae (Fig. 1). However, it is important to note that how DNA was mobilized from bacteria and inserted into the genomes of these nematodes in order to explain these HGT events is completely unknown (Haegeman et al. 2011). Interestingly, evidence for HGT in the other lineages of plant parasites, in particular the Longidoridae and Trichodoridae (Fig. 1), is either lacking or completely missing (Quist et al. 2015). This observation might suggest that HGT did not play a major role in the evolution of plant parasitism in these more basal lineages.

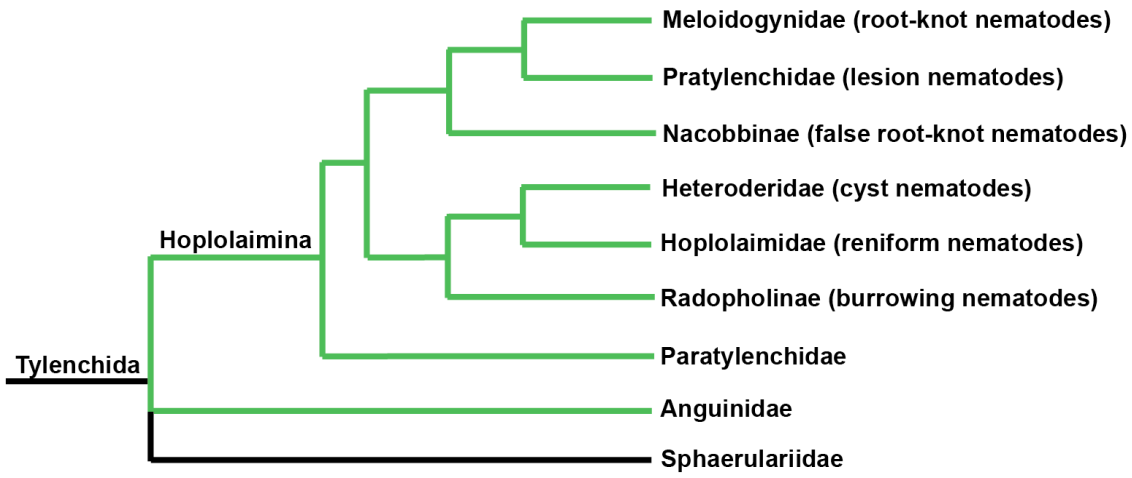



**Fig. 2. Cladogram of the order Tylenchida.** Green branches illustrate plant-parasitic lineages. Common names of the majority of Hoplolaimina plant parasites are indicated in parentheses. Cladogram is adapted from (Haegeman et al. 2011; Quist et al. 2015).

## 1.2 Agronomically important plant-parasitic nematodes

Plant-parasitic nematodes are among the most economically devastating plant pathogens causing an estimated US$100 billion in damages each year (Oka et al. 2000). The sedentary root endoparasites within Hoplolaimina (described in section 1.3) account for most of these damages (Jones et al. 2013). The majority of research on the sedentary root endoparasites focuses on those with the largest agronomic impact, the root-knot (*Meloidogyne* spp.; family Meloidogynidae) and cyst (i.e., *Heterodera* and *Globodera* spp.) nematodes (Fig. 2). Root-knot and cyst are the top two groups of economically important plant-parasitic nematodes and alone present a significant threat the global economy (Jones et al. 2013).

Though root-knot nematodes account for more total damages compared to cyst nematodes, this is mostly because they have a very wide host range infecting almost every vascular plant species (Jones et al. 2013). We therefore classify root-knot nematodes as generalists. Conversely, cyst nematodes are much more closely adapted to only a few or even a single host plant species (Lilley et al. 2011), and are thus classified as specialists. This feature of cyst nematodes often makes them of critical concern for a particular crop plant species as much damage can be inflicted at one time, and is experienced in most productive growing regions throughout the world. For example, the potato cyst nematodes (*Globodera pallida* and *G. rostochiensis*) are estimated to cause 9% in losses of total potato (*Solanum tuberosum*) production worldwide (Jones et al. 2013). Also, the soybean cyst nematode (*Heterodera glycines*) is the number one soybean

pathogen causing over US$1 billion in damages in the USA alone (Koenning and Wrather 2010).

The importance of plant-parasitic nematodes, in particular the root-knot and cyst nematodes, for the global economy and food security indicate the critical need for control strategies that will reduce their impact, and possibly eradicate them entirely. In the mid to late 1900s, harsh chemicals called nematicides (e.g., methyl bromide) were effective at reducing population levels of plant-parasitic nematodes in infested fields. However, presently, nematicide use is slated due to environmental problems and health concerns for animals and humans (Oka et al. 2000). Crop rotation has been a widely used method to prevent the build up of plant-parasitic nematode populations, in particular for cyst nematodes given their narrow host range (Oka et al. 2000). However, due to the ability of plant-parasitic nematodes to enter a prolonged period of dormancy, in particular the cyst nematodes [as long as twenty years (Jones et al. 2013)], crop rotation alone is often insufficient as a control strategy (Conley et al. 2011). Currently, the most success for reducing plant-parasitic nematode populations has been with natural genetic resistance, usually combined with crop rotation for cyst nematodes (Conley et al. 2011). However, due to massive population sizes [e.g., *H. glycines* population densities can reach greater than 30,000 eggs per $cm^3$ of soil (Chen et al. 2001)], host resistance is often overcome in only a few generations (Dong et al. 2005). Therefore, more robust control strategies must be developed against plant-parasitic nematodes in order to help sustain the global economy and food security, especially in the midst of a rapidly growing human population.

### 1.3 Infection process of sedentary endoparasitic nematodes

Most plant-parasitic nematodes infect roots, but where they infect within the roots, and for how long is quite variable. The approximately thirty day life cycles of the sedentary endoparasitic nematodes (e.g., root-knot and cyst nematodes) begin with embryonic development to the first-stage juveniles (J1s) inside of eggs in the soil. Still inside of the eggs, J1s molt to the pre-parasitic second-stage juveniles (pre-J2s). Cyst nematode pre-J2s hatch after sensing root exudates from their host plants, while root-knot nematode pre-J2s hatch based on soil temperature and moisture conditions regardless of sensing stimuli from root exudates, possibly in correlation with their wide host range. Pre-J2s penetrate the roots most often near the root tips and migrate either intercellularly through the apoplast (root-knot nematodes) or intracellularly (cyst nematodes) most often to the vascular cylinder. There, pre-J2s probe with their stylets for a vulnerable cell(s) to use for the formation of their feeding sites. Once found, the migratory pre-J2s switch to sedentary parasitic (par)-J2s and begin to form their feeding sites that will provide their nourishment for the remaining stages of their life cycles. Most often, feeding site formation continues through the second molt to the third-stage juveniles (J3s) and is completed before the third molt to the fourth-stage juveniles (J4s). Sex is most often determined in the J4 stage, and in the fourth and final molt to the adult stage the vermiform males regain their mobility and exit the roots. Most root-knot nematodes reproduce by parthenogenesis, and thus, males often have no role in reproduction. However, cyst nematodes reproduce by amphimixis, and thus, reemerged males search for the still sedentary pyriform females for copulation. Regardless of the mode of

reproduction, adult females can produce hundreds of eggs that are eventually left in the soil for the next generation (Hussey and Grundler 1998; Jones et al. 2013).

The elaborate feeding sites of root-knot and cyst nematodes are different in many ways. Root-knot nematodes select several vulnerable cells within the vascular cylinder, and sometimes cortex, and through repeated rounds of acytokinetic mitosis, these cells are transformed into highly metabolic 'giant cells'. Root-knot nematodes feed from either one of these giant cells during one life cycle. Cyst nematodes, however, select a single cell called the initial syncytial cell (ISC) usually of pericycle origin and initiate cell wall dissolution starting at the plasmodesmata. Cell wall dissolution spreads until numerous cells (sometimes in the hundreds) that surround the ISC are recruited to form a single multi-nucleated cell called the syncytium. Both giant cells and syncytia serve as metabolic sinks for compounds derived from photosynthesis and are vital for the survival of root-knot and cyst nematodes, respectively (Hussey and Grundler 1998; Jones et al. 2013).

## 1.4 Plant-parasitic nematode effector proteins

Another major adaptation of plant parasitism—especially in Hoplolaimina—was the evolution of elaborate unicellular secretory glands associated with the esophagus. The Hoplolaimina plant parasites contain one dorsal and two subventral esophageal glands. The subventral glands are positioned behind to the dorsal gland, but both gland types contain valves that connect to the esophageal lumen for transporting cargo to the stylet for release from the nematode. These glands produce proteins with N-terminal signal

peptides (called effectors) that are ultimately packaged into *trans*-Golgi-produced secretory granules that travel to the valves for exocytotic release into the esophageal lumen. The effectors then pass through the stylet for delivery into the infected host plant (Mitchum et al. 2013).

Molecular, biochemical and ultrastructural studies have shown that the esophageal glands are developmentally regulated, in both the timing of effector production and secretion, and in their morphological changes. In the J2 stage, the sizes of the subventral and dorsal glands are similar, but the majority of effector proteins at this stage appear to be produced and secreted from the subventral glands, though there are exceptions. However, in the subsequent stages (J3-adult) the subventral glands atrophy while the dorsal gland hypertrophies and appears to produce and secrete the majority of effectors during these stages. Accordingly, CWMPs such as the cellulases that assist nematode migration by loosening and degrading cell walls are produced and secreted from the subventral glands in pre-J2s. Moreover, effectors that contribute directly or indirectly to feeding site formation and/or maintenance are produced and secreted from the dorsal gland during the corresponding par-J2 to J4 stages, and sometimes all the way to the adult stage (Hussey and Grundler 1998; Davis et al. 2004; Mitchum et al. 2013).

Hogenhout et al. (2009) defined pathogen effectors as ‘all pathogen proteins that alter host cell structure and function’, and findings for most plant-parasitic nematode effectors are consistent with such a broad definition. Immunofluorescence and ultrastructural studies indicate that plant-parasitic nematode effectors are delivered into both the apoplast and cytoplasm of infected host plant cells. Molecular and biochemical studies indicate that in the apoplast, specific host proteins and cell wall components are

targeted for enhancing both migration and feeding site formation. Also, effectors in the cytoplasm of infected host cells target specific cellular compartments and host proteins for directly or indirectly contributing to the modulation of host cellular defenses and/or feeding site formation and/or maintenance. Although numerous studies have been published and are currently underway for functional characterization of effector proteins from plant-parasitic nematodes, we are only just beginning to understand the complexity of effector mechanisms. This information will be crucial in gaining a more thorough understanding of the complex interactions between nematodes and plants, and for developing novel control measures against these important plant pathogens (Hewezi and Baum 2013; Mitchum et al. 2013; Hewezi 2015).

### 1.5 Host responses to plant-parasitic nematodes

Histological studies have shown that in compatible (susceptible) interactions, root-knot and cyst nematodes induce dramatic alterations in feeding cells. These alterations include hypertrophied nuclei and nucleoli, complex cytoskeletal changes, increased ribosomes, reduced vacuoles, and extensive changes in cell wall architecture including cell wall ingrowths characteristic of transfer cells, among many other changes (Kim et al. 1986; Sobezak and Golinowski 2009; Mitchum et al. 2013). Also, transcriptomic and metabolomic studies indicate massive changes in gene expression [over 7,000 differentially expressed host genes in developing syncytia (Szakasits et al. 2009)] revealing a complex network during feeding site formation (Ithal et al. 2007b; Hewezi et al. 2012; Mitchum et al. 2013). In general, this novel gene expression network

involves the up-regulation of numerous genes involved in growth and developmental processes and the down-regulation of defense- and stress-related gene expression. Interestingly, there appears to be a complex coordination between growth and developmental pathways with stress and defense responses (Ithal et al. 2007b; Szakasits et al. 2009; Hewezi et al. 2012; Liu et al. 2014), but our current understanding is rudimentary. On the other hand, we do know that this coordination involves interplay between nematode effectors and host small RNA networks, as well as their targets within the host, as their genetic manipulation results in reduced susceptibility (Hewezi et al. 2008b; Hewezi et al. 2012; Hewezi and Baum 2013; Mitchum et al. 2013; Liu et al. 2014; Hewezi 2015; Hewezi and Baum 2015).

In resistant crop varieties, feedings site formation is interrupted and results in a hypersensitive cell death-like reaction with the ultimate death of the nematode juveniles (Kandoth and Mitchum 2013). Resistance in multiple crops is largely *Resistance* (*R*)-gene-mediated, indicating that the interactions between at least some nematodes and their host plants represent additional examples of the gene-for-gene model of host resistance (Kandoth and Mitchum 2013). However, the interaction between *H. glycines* and soybean appears to be a remarkable exception (Kandoth and Mitchum 2013). In this pathosystem, resistance is predominantly determined by two different loci, *Resistance to Heterodera glycines 1* and *4* (*Rhg1* and *Rhg4*) (Kandoth and Mitchum 2013). *Rhg1* contains tandem repeated copies of three different genes—*amino acid transporter*, *α-SNAP*, and *wound-inducible 12*—and the number of repeats determines the level of resistance (Cook et al. 2012; Cook et al. 2014). *Rhg4* contains a single gene—*serine hydroxymethyltransferase 1*—and resistance is allelic and determined by a single amino acid within the catalytic

domain of the encoded protein (Liu et al. 2012). Resistance to *H. glycines* can be exclusively *Rhg1*-mediated, or a combination of both *Rhg1* and *Rhg4* (Kandoth and Mitchum 2013). As this resistance is unique among plant-parasitic nematodes, and most other plant pathogens, this pathosystem represents an excellent model to study novel forms of resistance (Cook et al. 2012; Liu et al. 2012; Kandoth and Mitchum 2013; Cook et al. 2014).

## 1.6 General overview of research chapters

Due to the highly specialized and agronomically important interaction between *H. glycines* and soybean, the following research chapters greatly emphasize this pathosystem.

Effector proteins serve as probes for understanding the complex interactions between plant-parasitic nematodes and their host plants. Also, functional characterization of nematode effectors offers the potential to reveal vulnerable points in the parasitic cycle that may be targeted for potential control measures. Previous genomic and transcriptomic studies established relatively large lists of candidate effectors (referred to as parasitomes or effector repertoires) for both root-knot (Huang et al. 2003; Abad et al. 2008; Rutter et al. 2014) and cyst (Gao et al. 2003; Cotton et al. 2014; Thorpe et al. 2014) nematodes. However, predictions of complete parasitomes from the genome sequences of both root-knot (Abad et al. 2008) and cyst (Cotton et al. 2014) nematodes revealed that the current list of candidate effectors in *H. glycines* (Gao et al. 2001, 2003) is likely incomplete. Thus, the first research chapter (chapter 2) aims at identifying new candidate *H. glycines*

effectors that were missed in these previous gland-mining projects. In this work, we document eighteen new candidate effectors in *H. glycines* of which three are hypothesized to present novel HGT events.

The subsequent research chapter (chapter 3) builds on the work from chapter 2 and investigates the HGT hypotheses for the candidate *H. glycines* effectors in question. All three candidate effectors are concluded to have been part of more ancient HGT events from rhizosphere bacteria, but of course without an obvious mechanism. However, chapter 4 presents the first report of a possible mechanism for HGT in plant-parasitic nematodes, in particular for the Hoplolaimina plant parasites. This chapter reveals that a particular group of retroviruses that is specific to the distal clades of Nematoda is significantly associated with HGT genes in genomes of Hoplolaimina nematodes.

Also building on the work from chapter 2, chapters 5 and 6 investigate the function of one of the most highly expressed effectors found in the former study during *H. glycines* parasitism of soybean. The research reported in these chapters reveal numerous novelties for the *H. glycines*-soybean pathosystem, in particular that *H. glycines* targets an effector for efficient localization to plant cell nucleoli for suppression of innate immune responses. Furthermore, this chapter reveals that the *H. glycines* effector in question serves as an ideal target for the development of novel control measures in soybean.

Finally, chapter 7 presents a translational research project in soybean that was originally studied in the model cyst nematode interaction between the beet cyst nematode *Heterodera schachtii* and *Arabidopsis thaliana* (Hewezi et al. 2012). Briefly, the plant microRNA396-*Growth Regulating Factor* network delineates syncytium formation in

soybean, and interfering in the homeostasis of this network greatly reduces susceptibility to *H. glycines*. This chapter presents arguably one of the most attractive targets for engineering synthetic forms of resistance to *H. glycines*.

Collectively, these research chapters provide an extensive investigation of several aspects of plant parasitism by nematodes, with particular emphasis on *H. glycines*. This includes effector discovery and functional characterization, evolutionary analyses with particular emphasis on HGT, and host responses, including the formation of feeding sites and plant immunity to nematodes. This research will provide knowledge that will undoubtedly help to development novel control measures against plant-parasitic nematodes, and provide foundations for future studies of the complex interactions between nematodes and plants.

# CHAPTER 2. EIGHTEEN NEW CANDIDATE EFFECTORS OF THE PHYTONEMATODE *HETERODERA GLYCINES* PRODUCED SPECIFICALLY IN THE SECRETORY ESOPHAGEAL GLAND CELLS DURING PARASITISM

A paper published in *Phytopathology*

Jason B. Noon, Tarek Hewezi, Thomas R. Maier, Carl Simmons, Jun-Zhi Wei, Gusui Wu, Victor Llaca, Stephane Deschamps, Eric Davis, Melissa G. Mitchum, Richard Hussey, Thomas J. Baum

Noon JB, Hewezi T, Maier TR, Simmons C, Wei JZ, Wu G, Llaca V, Deschamps S, Davis EL, Mitchum MG, Hussey RS, Baum TJ. 2015. Eighteen new candidate effectors of the phytonematode *Heterodera glycines* produced specifically in the secretory esophageal gland cells during parasitism. Phytopathology. 105:1362-1372.

## Abstract

*Heterodera glycines*, the soybean cyst nematode, is the number one pathogen of soybean (*Glycine max*). This nematode infects soybean roots and forms an elaborate feeding site in the vascular cylinder. *H. glycines* produces an arsenal of effector proteins in the secretory esophageal gland cells. More than 60 *H. glycines* candidate effectors were identified in previous gland-cell-mining projects. However, it is likely that additional candidate effectors remained unidentified. With the goal of identifying remaining *H. glycines* candidate effectors, we constructed and sequenced a large gland

cell cDNA library resulting in 11,814 expressed sequence tags. After bioinformatic filtering for candidate effectors using a number of criteria, *in situ* hybridizations were performed in *H. glycines* whole-mount specimens to identify candidate effectors whose mRNA exclusively accumulated in the esophageal gland cells, which is a hallmark of many nematode effectors. This approach resulted in the identification of 18 new *H. glycines* esophageal gland-cell-specific candidate effectors. Of these candidate effectors, 11 sequences were pioneers without similarities to known proteins while 7 sequences had similarities to functionally annotated proteins in databases. These putative homologies provided the bases for the development of hypotheses about potential functions in the parasitism process.

## 2.1 Introduction

*Heterodera glycines*, the soybean cyst nematode, is a root-infecting plant-parasitic nematode of the subfamily Heteroderinae (Subbotin et al. 2010) and is the most serious pathogen of soybean (*Glycine max*), causing estimated annual U.S. soybean yield losses of more than one billion dollars (Koenning and Wrather 2010; Conley et al. 2011; Liu et al. 2012). *H. glycines* undergoes postembryonic development to the second-stage juvenile (J2) inside its egg and then hatches. Hatched pre-parasitic J2 (pre-J2) become parasitic (par-J2) when they penetrate soybean roots and migrate intracellularly through cortical parenchyma to the vascular cylinder where they become sedentary and select an initial feeding cell. Effector proteins secreted into the initial feeding cell suppress plant defenses and initiate the redifferentiation of this cell into a syncytial feeding site composed of

numerous fused cells that have undergone cell wall dissolution, resulting in a continuous multinucleated cytoplasm (syncytium) (Hussey and Grundler 1998). The syncytium provides the essential source of nourishment for the growth and development of the nematode into subsequent sedentary life stages. While the nematodes complete their postembryonic development through two additional juvenile stages (J3 and J4) into adult males and females, continued delivery of effector proteins modulates effective suppression of plant defenses and the full development of the syncytium while also mediating the feeding process.

Major factors contributing to the evolutionary success of phytonematodes like *H. glycines* are their adaptations to plant parasitism (Hussey 1989; Baldwin et al. 2004; Davis et al. 2004; Mitchum et al. 2013), which include large specialized secretory gland cells (one dorsal and two subventral) associated with the esophagus. These esophageal glands produce nematode effectors as secretory proteins that are delivered through the hollow nematode mouth spear (stylet) into the plant during parasitism (Hussey 1989; Mitchum et al. 2013).

Most known phytonematode effectors are encoded by genes that are expressed exclusively in the three esophageal gland cells (Davis et al. 2000; Mitchum et al. 2013). The dual subventral gland cells are most active during the early stages of parasitism (i.e., during migration and the early events of syncytium formation) while the single dorsal gland cell is most active during and following the onset of syncytium formation, which is accompanied by hypertrophy of the dorsal gland cell and ultimately atrophy of the subventral gland cells (Hussey and Grundler 1998; Davis et al. 2004).

Small suites of parasitism genes that encode phytonematode effectors were first cloned through differential gene expression and target gene identification (Ding et al. 1998; Smant et al. 1998; Lambert et al. 1999; Qin et al. 2000). In the first relatively high-throughput approach towards obtaining comprehensive cohorts of phytonematode candidate effectors (i.e., the parasitome), the contents of the esophageal gland cells were microaspirated at multiple parasitic stages, and the isolated mRNA was used to construct cDNA libraries enriched for candidate nematode effectors (Gao et al. 2001; Wang et al. 2001; Gao et al. 2003; Huang et al. 2003). Individual clones were sequenced and bioinformatically analyzed for the presence of N-terminal signal peptides (Petersen et al. 2011) and the absence of transmembrane (TM) domains (Sonnhammer et al. 1998; Krogh et al. 2001). Spatial expression of candidate effectors was then assessed by *in situ* mRNA hybridization (de Boer et al. 1998) in nematode specimens, and approximately fifty new *H. glycines* candidate effectors at that time were identified as exclusively expressed in the esophageal glands (Gao et al. 2003). Interestingly, around 75% of these candidate effectors were considered 'pioneer' sequences without similarities to any sequences in databases at that time.

Since then, several *H. glycines* candidate effectors have been functionally characterized and have been found to exhibit specific subcellular localizations within plant cells (Elling et al. 2007; Hewezi and Baum 2013) and to interact with host plant proteins for promoting successful parasitism (Hewezi and Baum 2013; Mitchum et al. 2013), and thus were determined to be *bona fide* effectors. Functional classifications of cyst nematode effectors have revealed complex cell wall modifications either for enhancing migration or for syncytium formation (Smant et al. 1998; Gao et al. 2001;

Vanholme et al. 2007; Hewezi et al. 2008a), developmental cellular reprogramming through mimicry of plant peptide hormones (Wang et al. 2005; Guo et al. 2011; Replogle et al. 2011; Wang et al. 2011) and modulation of phytohormone transport (Lee et al. 2011), and suppression and activation of plant defense responses either through direct interaction with host plant immune regulators (Hamamouch et al. 2012; Lozano-Torres et al. 2012; Postma et al. 2012) or indirectly through interaction with host plant targets that are associated with defense signaling pathways (Hewezi et al. 2010; Patel et al. 2010; Chronis et al. 2013).

Phytonematode effector characterization provides the potential to develop novel control measures through the manipulation of vulnerable points identified in the parasitic cycle (Davis et al. 2004; Hewezi and Baum 2013; Mitchum et al. 2013). Since the repertoire of candidate effectors in the potato cyst nematode (*Globodera pallida*) predicted from a whole genome sequence (Cotton et al. 2014; Thorpe et al. 2014) is considerably larger than the cohort obtained from expression analyses (Smant et al. 1998) it is likely that the current known repertoire of candidate effectors in *H. glycines* (Mitchum et al. 2013) is incomplete. In the current absence of a published robust genome sequence of *H. glycines*, and with the goal of identifying additional novel *H. glycines* candidate effectors that were missed in prior gland-mining projects, a new and significantly larger cDNA library was prepared from the gland region of *H. glycines* similar to the work described by Gao and colleagues (Gao et al. 2003), but with minor modifications as described below.

In this current report, we identified eighteen new *H. glycines* candidate effectors that are expressed specifically in the esophageal gland cells, lack canonical endoplasmic

reticulum (ER) retention signals, and contain predicted N-terminal signal peptides for likely secretion from the glands and into the plant. While more than half of these eighteen new candidate effectors were found to be pioneer sequences, some showed detectable sequence similarities, which suggested horizontal gene transfer as a possible origin of a few of these sequences in the nematode genome. Furthermore, these similarities also allowed the formulation of biological hypotheses about their function. These candidate effectors include several enzymes that may be involved in the suppression of host cellular defenses, and some that may help weaken physical barriers to infection. Also, some of these new candidate effectors may assist in the formation and function of the syncytium through pathways not previously considered.

## 2.2 Materials and Methods

### 2.2.1 Sequence data

Nucleotide and amino acid sequence data have been submitted to the DDBJ/EMBL/GenBank databases. All generated expressed sequence tags (ESTs) are available in dbEST under the accessions JZ682331-JZ693590, library accession number LIBEST_028433.

### 2.2.2 *H. glycines* gland cell mRNA purification and cDNA amplification

Parasitic stages of *H. glycines* inbred line OP50 (Fig. 1A) were hand-dissected from infected soybean roots, surface-sterilized, and embedded in 0.7% agarose. Esophageal gland cell cytoplasm was microaspirated from 100 *H. glycines* specimens

(mixed stages) with glass micropipettes containing 1μL of mRNA extraction buffer (Gao et al. 2001) and transferred into microcentrifuge tubes for storage at -80˚C (Fig. 1A).

Oligo $(dT)^{25}$ magnetic beads (DynaI, Lake Success, NY, U.S.A.) were used to isolate mRNA from the aspirated gland-cell cytoplasm (Fig. 1A). The isolated mRNA was then eluted with 5:1 diethyl pyrocarbonate (DEPC)-treated $H_2O$ at 70˚C for 2 min (Gao et al. 2001). First-strand cDNA synthesis was performed with the Super SMART cDNA Synthesis Kit (Clontech Laboratories, Palo Alto, CA, U.S.A.) according to the manufacturer's instructions (Fig. 1A). Mixtures were then diluted with 90μL TE buffer (10mM Tris, pH 7.6, and 1 mM EDTA). Long-distance (LD) PCR was then performed with 10μL first-strand reaction solution, 2μL 10mM dNTP mix, 10μL TaqPlus long 10x low salt buffer, 1μL TaqPlus long (Stratagene, La Jolla, CA, U.S.A), and 2μL nested universal primer (Clontech) (Fig. 1A). LD PCR consisted of hot start followed by an optimum 27 cycles at 94˚C for 20s, 65˚C for 30s, and 72˚C for 7min. DEPC-treated $H_2O$ was used as a negative control at each reaction step above.

### 2.2.3 Gland-cell cDNA library preparation

The QIAquick PCR purification kit (Qiagen, Valencia, CA, U.S.A.) was used to purify the LD PCR-amplified gland cell cDNA (Gao et al. 2001). Then, the purified cDNA was ligated into pGEM-T Easy (Promega, Madison, WI, U.S.A.) at a 3:1 mass ratio (cDNA/plasmid) at 4˚C overnight (Fig. 1A). The ligated products were then precipitated with 10mM glycogen and 100% ethanol, and then washed with 70% ethanol. Purified ligation products were then electroporated into competent *E. coli* XL10-GOLD cells for blue-white selection (Fig. 1A). We handpicked as many white colonies as

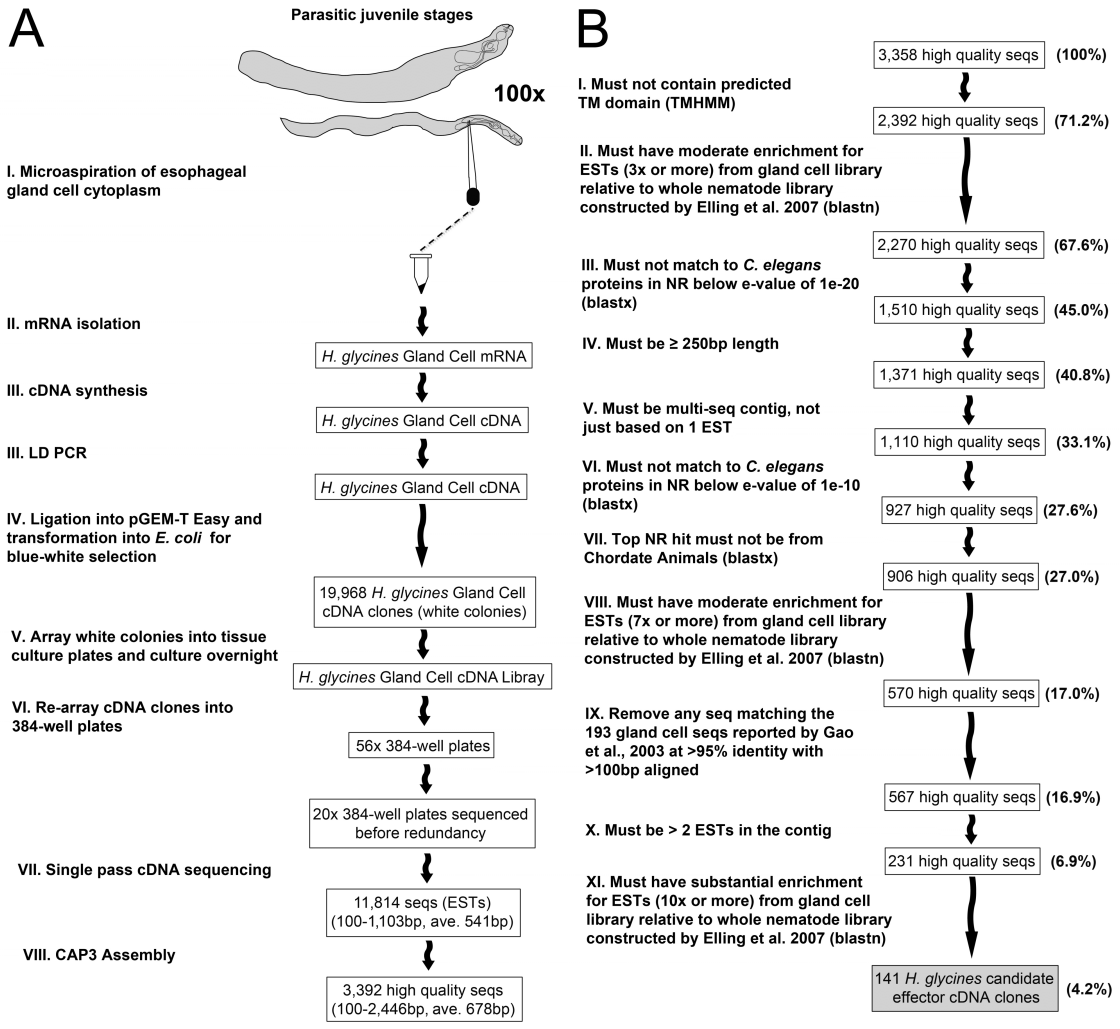



**Fig. 1. Overview of methodology. (A)** Preparation and sequencing of the *H. glycines* gland cell cDNA library. **(B)** Bioinformatic pipeline for enrichment of candidate *H. glycines* effectors.

possible, and transferred them to 96-well Microtest III tissue culture plates (Becton Dickinson, Franklin Lakes, NJ, U.S.A.) with 200µL Luria-Bertani medium, 10% glycerol and ampicillin (100µg/mL), and incubated overnight at 37˚C (Fig. 1A).

Quality control of the cDNA library was assessed using two methods. First, PCR was performed using the cDNA as template to test if the four previously identified *H. glycines* candidate effectors 2B10, 3B05, 30D08, and 25G01 could be amplified. Second,

EcoRI restriction digests were performed on 24 gland-cell cDNA library clones in order to evaluate insert sizes, empty pGEM-T Easy vector was used as a negative control. From the first test, all four candidate effectors were successfully amplified, and from the second, inserts were found for all 24 clones tested that fell into 2 size categories, 12 clones were from 0.4-1.0-Kbp in size and the other 12 clones were from 1.0-2.4-Kbp. Therefore, the quality of the gland-cell cDNA library was determined to be ideal for sequencing.

### 2.2.4 cDNA sequencing and assembly

The cDNA clones were then re-arrayed to 384-well plates (Fig. 1A). Plates were randomly selected for single pass cDNA sequencing using the SMART forward primer 5'-AAGCAGTGGTATCAACGCAGAGTACGCG-3' and an oligo (dT) reverse primer with equimolar amount of $(T)_{21}A$, $(T)_{21}C$ and $(T)_{21}G$ until sequences became redundant, (Fig. 1A). Sequences were collected on an ABI 3700 Sequencer (Applied Biosystems, Foster City, CA, U.S.A.). Base calls and quality scores were generated from the raw chromotograph files using Phred (Ewing and Green 1998; Ewing et al. 1998). Our own script was used (seqclean) for primary vector-linker cleanup, and for removing low-quality sequences. The CAP3 program was used to assemble the ESTs into high quality contiguous sequences (Fig. 1A).

### 2.2.5 Bioinformatic analyses for enrichment of candidate *H. glycines* effectors

The high quality sequences that resulted from the CAP3 program were evaluated using the Spotfire DecisionSite functional genomics software (Kaushal and Naeve 2004).

First, we searched for candidate effectors reported by Gao and colleagues (Gao et al. 2003) using blastn and removed any sequence that resulted in greater than 95% identity with greater than 100 aligned-bp. Then, we generated a bioinformatic pipeline to enrich for candidate *H. glycines* effector cDNAs, which consisted of eleven filtering criteria (see Fig. 1B for a complete, step-by-step overview). Note that we progressively increased the stringency of our pipeline in order to allow us to better evaluate the kinds of sequences that survived at each step, rather than elimating them all at once, which also allowed to better evalute the quality of our sequence collection. In the first step, the high quality sequences were evaluated for the presence of predicted TM domains using the TMHMM server (Sonnhammer et al. 1998; Krogh et al. 2001). All sequences that contained predicted TM domains while not simultaneously containing predicted signal peptides using SignalP4.0 (Petersen et al. 2011) were removed. In the second step, sequences were evaluated via blastn for having 3x or more relative enrichment for ESTs from the gland cell library compared to the *H. glycines* whole nematode library constructed by Elling and colleagues (Elling et al. 2009). All sequences that were not enriched 3x or more for the gland cells relative to whole nematodes were removed. The third step removed all sequences that matched to *Caenorhabditis elegans* proteins in the non-redundant database (NR) below an E-value of 1E-20 using blastx. In the fourth step, all sequences were removed if they were less than 250-bp in length. The fifth step removed all singletons (unassembled sequences – made up of only 1 EST). In the sixth step, all sequences that matched to *C. elegans* proteins in NR below an E-value of 1E-10 using blastx were removed. In the latter step, we used such a high stringency to largely avoid testing any sequence that contained significant similarity to sequences from a free-living ancestor,

thereby increasing our confidence of the candidate sequences encoding effectors rather than proteins common to non-parasitic organisms. Although additional phytonematode effectors may contain domains shared with *C. elegans* proteins, we chose not to allow for such flexibility, as this would probably have largely increased the number of false positives that would have made it through our pipeline. Using blastx against NR, step seven removed any sequence that retrieved a protein from chordate animals as the best match. In step eight, sequences were evaluated via blastn for having 7x or more relative enrichment for ESTs from the gland cell library compared to the *H. glycines* whole nematode library constructed by Elling and colleagues (Elling et al. 2009). Step nine removed any sequence that matched to the 193 *H. glycines* gland cell cDNA sequences obtained by Gao and colleagues (Gao et al. 2003) at greater than 95% identity and with greater than 100-bp aligned using blastn. In step ten, sequences were removed if they did not contain more than two ESTs in the contig. Finally, in step eleven, sequences were evaluated via blastn for having 10x or more relative enrichment for ESTs from the gland cell library compared to the *H. glycines* whole nematode library constructed by Elling and colleagues (Elling et al. 2009).

### 2.2.6 Developmental expression of candidate effectors

For all 141 gland-cell cDNA clones that passed the bioinformatic filtering (Fig. 1B), specific forward and reverse primers were used to synthesize digoxigenin (DIG)-labeled sense and antisense DNA probes (Boehringer Mannheim, Germany) by asymmetric PCR (Gao et al. 2001). *In situ* hybridizations were performed on fixed, permeabilized preparasitic and mixed parasitic *H. glycines* stages (de Boer et al. 1998;

Gao et al. 2001). Alkaline phosphatase-conjugated anti-DIG antibody and substrate were used to detect probes that hybridized within the nematode specimens (de Boer et al. 1998). The specimens were observed with a Zeiss Axiovert inverted compound light microscope.

### 2.2.7 Gene model predictions

We performed blastn searches with the nucleotide sequences from all gland-positive cDNA clones against a draft *H. glycines* genome sequence (Matt Hudson and Kris Lambert, personal communication) with an E-value cutoff of 1E-10. For each clone, the sequence within the scaffold that aligned at nearly 100% identity, including the 5'- and 3'-flanking sequences, were submitted to the self-training eukaryote gene prediction software GeneMark.hmm (Lomsadze et al. 2005) using the test set from the *C. elegans* genome. For each of the resulting gene models the exon sequences were combined and translated with the ExPASy Translate tool to obtain the resulting putative full-length protein sequences. Finally, for predicting the N-terminal signal peptides we used the software SignalP4.0 with the following parameters: Organism group=Eukaryotes, default D-cutoff value=0.50 for SignalP-TM networks, method=input sequences may include TM regions.

### 2.2.8 Protein sequence similarity searches and domain analyses

To search for putative homologs of the identified *H. glycines* candidate effectors, blastp and PSI-blastp (Altschul et al. 1997) searches were performed against NR with E-value cutoffs equal to 0.001. To search for conserved protein domains, both blastp with

the conserved domains database CDD v3.11 – 45746 position-specific scoring matrices (PSSMs) (Marchler-Bauer et al. 2011) and InterProScan (Quevillon et al. 2005) were used with default parameters.

## 2.3 Results

### 2.3.1 Sequencing and bioinformatic analysis of an *H. glycines* gland cell cDNA library

We constructed a cDNA library from the esophageal gland cells of *H. glycines* that contained 19,968 plasmid cDNA clones (Fig. 1A). As detailed in the Materials and Methods section, our sequencing generated 11,814 sequences (i.e., ESTs), of 100- to 1,103-bp in length and an average length of 541-bp (Fig. 1A). Sequence assembly produced 3,392 distinct high quality cDNA sequences with an average length of 678-bp and totaling 2.3-Mbp of unique sequence (Fig. 1A).

In order to assess the quality of this gland cell library, we searched the list of 3,392 high quality sequences for the candidate effectors reported by Gao and colleagues (Gao et al. 2003). We identified sequences representative of 34 of the 53 candidate effectors (64%) that were reported by Gao and colleagues (Gao et al. 2003) (Table S1, Appendix A), all of which were removed from the list, which reduced the number of high quality sequences to 3,358. The absence of 19 effectors (36%), all of which were pioneers, could be due to either technical differences such as different sampling times or different extraction and processing details or differences between the two *H. glycines* populations used. Nonetheless, having rediscovered the majority of the candidate

effectors reported by Gao and colleagues (Gao et al. 2003) was indicative that the quality of the gland cell library and sequencing was suitable for downstream analyses.

In all, 11 bioinformatic filtering criteria were established in order to eliminate sequences that were unlikely to be gland-specific in expression and to enrich for possible candidate effectors (Fig. 1B). This set of filtering criteria reduced the 3,358 high quality sequences to 141 (4.2%). These retained sequences all fulfilled the following criteria: (i) comprised of more than two ESTs, (ii) different from the gland cell sequences generated by Gao and colleagues (Gao et al. 2003), (iii) unique from protein sequences of *C. elegans* and chordate animals, and (iv) without TM domains (Fig. 1B).

Gao and colleagues (Gao et al. 2003) performed nematode *in situ* hybridization screens on a total of 193 cDNA clones, of which 140 did not localize to the gland cells. As a testimony to the stringency of our filtering pipeline, our sequences retained after filtering (steps one through eight) only contained 3 of these 140 non-gland cell-specific sequences (Fig. 1B; step nine). In other words, this result indicated that the high stringencies of steps one through eight of our filtering pipeline was successful at removing non-gland cell-specific sequences.

### 2.3.2 Screening for gland cell-specific mRNA accumulation

To test for gland cell-specific mRNA expression in *H. glycines* specimens, we used DIG-labeled sense and antisense DNA probes from all 141 sequences in whole-mount nematode *in situ* mRNA hybridization tests. These analyses resulted in the identification of 18 distinct cDNA clones whose antisense probes hybridized to mRNA transcripts accumulating within the subventral (2 clones) or dorsal (16 clones) gland cells

of *H. glycines* (Fig. 2; Table 1). We designated these 18 sequences as *GLAND1* through *18*. All remaining cDNA clones either hybridized to tissues or cells other than the esophageal glands, showed non-specific hybridization patterns, or could not be localized in *H. glycines* specimens. None of the negative control sense probes showed any hybridization within *H. glycines* specimens (Fig. 2S).

Of the 18 gland-positive cDNA clones identified, nine clones were missing the complete N-termini of their encoded proteins, and could not be scrutinized for the presence of a predicted N-terminal signal peptide for secretion. However, we were able to access an unpublished *H. glycines* draft genome (Matt Hudson and Kris Lambert, personal communication) to search for all 18 gland-positive cDNAs in order to identify complete full-length coding sequences. High scoring scaffolds were identified from the draft genome sequence for all 18 gland-positive cDNA clones. Gene model predictions resulted in the identification of putative full-length protein sequences for all 18 clones. Importantly, we identified at least one intron in 17 of the 18 genes (Fig. 3). The only gene that was not found to contain any introns was *GLAND9* (Fig. 3), which was found to be by far the smallest of the GLAND proteins encoding a peptide of only 94-aa (Table 1). Thus, these findings, in addition to the observed gland cell-specific expressions (Fig. 2), indicated that the 18 sequences are encoded by the nematode genome and not contaminants.

Subsequently, all 18 candidates were confirmed to encode predicted N-terminal signal peptides, including the nine clones that were originally missing their complete N-termini of the encoded proteins (Fig. 3; Table 1). We next searched the protein sequences of all 18 candidates for the presence of a C-terminal animal (KDEL)- or yeast (HDEL)-

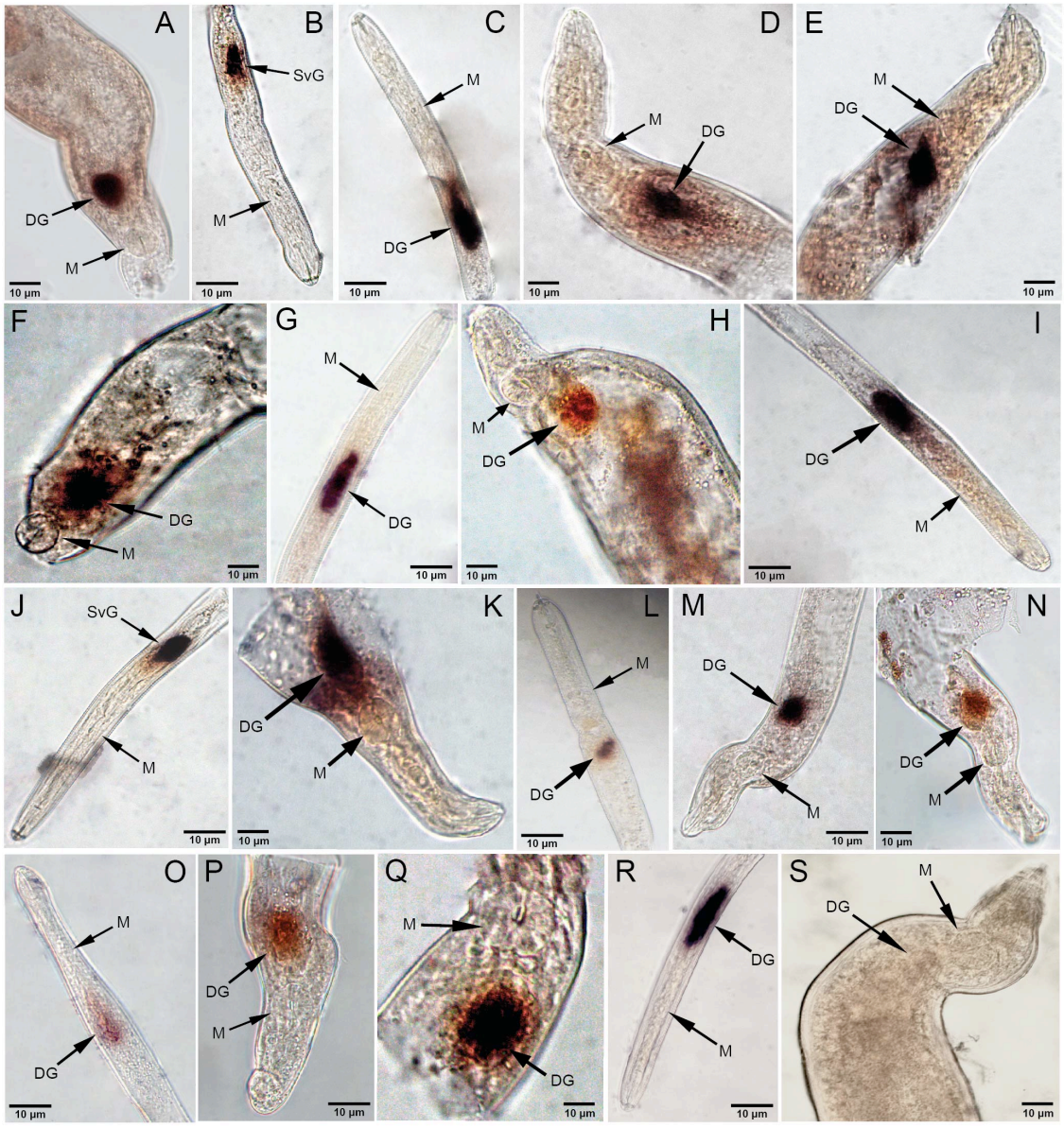



**Fig. 2. Hybridization of digoxygenin-labeled antisense DNA probes (dark staining) of the GLAND protein cDNAs to transcripts expressed exclusively within the subventral or dorsal esophageal gland cells of *Heterodera glycines.* (A)** GLAND1: parasitic third-stage juvenile. **(B)** GLAND2: migratory pre-parasitic second-stage juvenile. **(C)** GLAND3: migratory pre-parasitic second-stage juvenile. **(D)** GLAND4: parasitic third-stage juvenile. **(E)** GLAND5: parasitic third-stage juvenile. **(F)** GLAND6: parasitic third-stage juvenile. **(G)** GLAND7: parasitic second-stage juvenile. **(H)** GLAND8: parasitic third-stage juvenile. **(I)** GLAND9: parasitic second-stage juvenile. **(J)** GLAND10: migratory pre-parasitic second-stage juvenile. **(K)** GLAND11: parasitic third-stage juvenile. **(L)** GLAND12: parasitic second-stage juvenile. **(M)** GLAND13: parasitic third-stage juvenile. **(N)** GLAND14: parasitic third-stage juvenile. **(O)** GLAND15: parasitic second-stage juvenile. **(P)** GLAND16: parasitic third-stage juvenile. **(Q)** GLAND17: parasitic third-stage juvenile. **(R)** GLAND18: parasitic second-stage

juvenile. **(S)** Negative control sense probe of GLAND18: parasitic third-stage juvenile. DG = dorsal gland cell, M = metacorpus, SvG = subventral gland cells. Scale bars equal 10 μm.

**Table 1. Summary of the 18 GLAND proteins preceded by signal peptides for secretion and whose mRNAs are expressed exclusively in the esophageal gland cells of *Heterodera glycines***

| | | | | | | Gland expression[e] | | |
|---|---|---|---|---|---|---|---|---|
| **Clone** | **Accession[a]** | **Signal Peptide[b]** | **Protein[c]** | **Highest protein similarity[d]** | **E-value** | **Pre-J2** | **Par-J2** | **J3-A** |
| GLAND1 | KJ825712 | Yes - D=0.787[f] | 419 | GNAT – *Streptomyces** | 1E-21 | _[g] | DG | DG |
| GLAND2 | KJ825713 | Yes - D=0.934 | 234 | Pioneer | | SvG | SvG | SvG |
| GLAND3 | KJ825714 | Yes - D=0.732 | 460 | G12H04 – *H. glycines* | 8E-87 | DG | DG | DG |
| GLAND4 | KJ825715 | Yes - D=0.866 | 167 | 1106_3E10 – *G. rostochiensis* | 3E-26 | _ | DG | DG |
| GLAND5 | KJ825716 | Yes - D=0.861 | 187 | G11A06 – *H. glycines* | 4E-96 | _ | _ | DG |
| GLAND6 | KJ825717 | Yes - D=0.846 | 203 | 4D06 – *H. glycines* | 2E-104 | _ | _ | DG |
| GLAND7 | KJ825718 | Yes - D=0.681 | 367 | G15A10 – *H. glycines* | 3E-116 | _ | DG | _ |
| GLAND8 | KJ825719 | Yes - D=0.810 | 224 | Pioneer | | _ | _ | DG |
| GLAND9 | KJ825720 | Yes - D=0.837 | 94 | Pioneer | | _ | DG | _ |
| GLAND10 | KJ825721 | Yes - D=0.787 | 155 | CBP – *H. schachtii* | 4E-09 | SvG | _ | _ |
| GLAND11 | KJ825722 | Yes - D=0.810 | 611 | Pioneer | | _ | _ | DG |
| GLAND12 | KJ825723 | Yes - D=0.541 | 201 | Pioneer | | _ | DG | DG |
| GLAND13 | KJ825724 | Yes - D=0.618 | 595 | Invertase – *Rhizobium** | 8E-37 | _ | DG | DG |
| GLAND14 | KJ825725 | Yes - D=0.634 | 170 | Endopeptidase – *A. suum** | 7E-55 | _ | _ | DG |
| GLAND15 | KJ825726 | Yes - D= 0.837 | 496 | G23G11 – *H. glycines** | 4E-05 | _ | DG | DG |
| GLAND16 | KJ825727 | Yes - D=0.668 | 1,149 | CM – *H. glycines** | 4E-23 | _ | DG | DG |
| GLAND17 | KJ825728 | Yes - D=0.752 | 269 | DUO-3 – *C. elegans** | 2E-04 | _ | _ | DG |
| GLAND18 | KJ825729 | Yes - D=0.834 | 168 | Pioneer | | DG | DG | DG |

[a] Sequences submitted to GenBank.

[b] Signal peptides were predicted with the SignalP 4.0 software.

[c] Listed are the number of amino acids in the predicted protein sequences.

[d] GLAND proteins listed with asterisks were found to contain domains from either InterProScan or the Conserved Domains Database.

[e] *In situ* hybridization of antisense DNA probes to mRNA specifically within the dorsal esophageal gland cell (DG) or subventral esophageal gland cells (SvG) in pre-parasitic second-stage juveniles (Pre-J2), parasitic J2 (Par-J2), or parasitic J3, J4 or young adult stages (J3-A) of *Heterodera glycines*.

[f] The D-value cutoff was set to 0.5.

[g] Not detected.

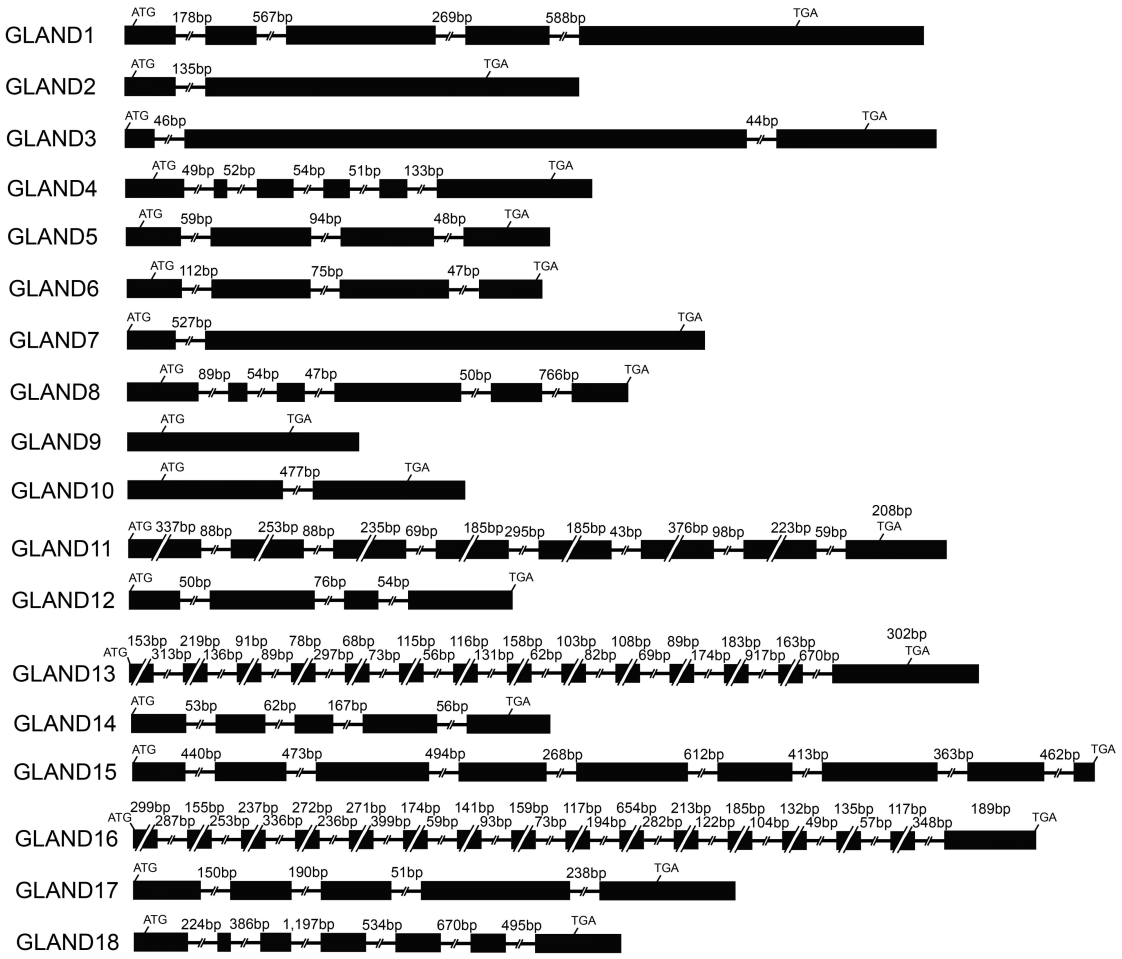



**Fig. 3. Gene models for *Hg-GLAND1* to *18*.** Gene model predictions were made for the sequences within the draft *H. glycines* genome that aligned with highest similarities to *Hg-GLAND1* to *18*. Exons are illustrated as black boxes and introns as horizontal lines. All introns are drawn with equal sizes for ease of presentation – the actual sizes are shown above. Exons for *GLAND11, 13* and *16*, except for the last exons for accurate placement of the stop codons, are also drawn equal in size within each gene for ease of presentation as these genes are much larger than the rest – the actual exon sizes are also shown above. Start and stop codons are also indicated above the respective exons for each gene.

type ER retention signal and for possible ER retrieval signals matching the N-terminal XXRR and C-terminal KKXX motifs with PSORT II (Nakai and Horton 1999). None were predicted to contain an ER retention signal, and GLAND12 was the only candidate predicted to contain a possible ER retrieval signal matching the KKXX motif, KKRA at

the C-terminus. However, since predicted ER retrieval signals (XXRR and KKXX) are neither necessary nor sufficient for localization of proteins to the ER membrane, there is no significance associated with this prediction (Nakai and Horton 1999). Thus, the lack of ER retention signals indicated that the 18 candidate proteins are unlikely localized in the ER of the gland cells. Furthermore, since Golgi-resident proteins are dependent on the properties of TM domains (Banfield 2011), of which none of the 18 candidates were found to contain, the encoded proteins are unlikely localized to the Golgi of the gland cells. Thus, we have identified 18 new gland-specific cDNA clones that encode secretory peptides that do not contain TM domains and are likely secreted from the gland cells. GLAND1 through 18 can be regarded as new *H. glycines* candidate effectors (Table 1).

### 2.3.3 Gland cell expression throughout *H. glycines* development

The timing of candidate effector gene expression during the life cycle frequently is informative when trying to infer effector function. We therefore qualitatively assessed *in situ* hybridization profiles of the 18 new candidate effectors in different developmental stages (Table 1). Among the 18 candidate effectors, three were most active during migratory (*GLAND10*) and early parasitic J2 stages (*GLAND7* and *9*). Interestingly, only one of these three was expressed in the subventral gland cells (*GLAND10*), although this cell type is typically most active during these early time points. Eleven candidate effectors were most active during sedentary parasitic stages (*GLAND1*, *4-6*, *8*, and *11-17*) and, as expected, all were expressed in the dorsal gland cell. The remaining three candidate effectors (*GLAND2*, *3* and *18*) were highly active throughout the entire life cycle. Of these, one was expressed in subventral gland cells (*GLAND2*) whereas

*GLAND3* and *18* were localized in the dorsal gland cell. These data suggest that GLAND7, 9 and 10 likely function during the migratory phase or the early sedentary phase of syncytium induction. On the other hand, GLAND1, 4-6, 8, and 11-17 likely function during syncytium formation and/or feeding. Finally, GLAND2, 3 and 18 appear important throughout the parasitic life cycle of *H. glycines*.

### 2.3.4 GLAND proteins with putative homologs in databases

To search for putative non-phytonematode homologs for GLAND proteins 1-18, blastp was performed against NR. Six GLAND proteins (GLAND2, 8, 9, 11, 12, and 18) did not match any known sequences below the cutoff E-value of 0.001, and thus, were immediately designated as pioneers (Table 1). Eight GLAND proteins matched only to other cyst nematode sequences. The top hit for GLAND3 was the previously identified *H. glycines* candidate effector 12H04 (GenBank accession AAO85452.1) with an E-value of 8.00E-87. However, large portions of the GLAND3 sequence do not align to 12H04, thus, the similarity is only partial. The top hit (E-value of 3.00E-26) for GLAND4 was the 3E10 isoform of the potato cyst nematode *Globodera rostochiensis* 1106 candidate effector family (GenBank accession AFH68219.1). GLAND5 hit to the previously identified *H. glycines* candidate effector 11A06 (GenBank accession AAP30754.1) with an E-value of 4.00E-96, while the top hit (E-value of 2.00E-104) for GLAND6 was the previously identified *H. glycines* candidate effector 4D06 (GenBank accession AAN32892.1). Although the sequence similarities between GLAND5 and 11A06 and GLAND6 and 4D06 are high, there are significant amino acid differences that distinguish these protein pairs (16% and 12% different amino acids, respectively). Hence, GLAND5

and GLAND6 are likely within candidate effector families containing 11A06 and 4D06, respectively, but they may have completely different functions or host targets. The top hit (E-value of 3.00E-116) for GLAND7 was the previously identified putative *H. glycines* gland protein G15A10 (GenBank accession AAP30765.1). Although the E-value is very low, much of the GLAND7 sequence is unique and does not align with G15A10. The G15A10 was not tested by *in situ* hybridization when initially identified (Gao et al. 2003) because the missing 5'-sequence prevented the prediction of a signal peptide. Given our new data with GLAND7, it now is likely that G15A10 also is an *H. glycines* candidate effector. The top hit (E-value of 4.00E-09) for GLAND10 was the functionally characterized cellulose-binding protein effector of *H. schachtii* (Genbank accession ABY49997.1) (Gao et al. 2004; Hewezi et al. 2008a), however, 67% of the amino acids are different. GLAND15 hit to another previously identified putative *H. glycines* gland protein G23G11 (GenBank accession AAP30771.1; E-value of 4.00E-05). Similar to G15A10, G23G11 also was not tested by *in situ* hybridization due to a missing 5'-end sequence, and thus, G23G11 also likely is an *H. glycines* candidate effector. The top hit for GLAND16 was the previously identified chorismate mutase candidate effector of *H. glycines* (GenBank accession AAO19577.2; E-value of 4.00E-23); however, GLAND16 is over four times larger. Four GLAND proteins hit to non-phytonematode sequences. The top hit for GLAND1 was a GCN5-related N-acetyltransferase from *Streptomyces* (GenBank accession WP_030417594.1) with an E-value of 1.00E-21. The top hit for GLAND13 was a beta-fructofuranosidase of *Rhizobium leguminosarum* (GenBank accession WP_003572067.1) with an E-value of 8.00E-37. GLAND14 hit to a prolyl endopeptidase of *Ascaris suum* (GenBank accession ERG83141.1) with an E-value of

7.00E-55. Finally, GLAND17 hit to isoform C of Protein DUO-3 from *C. elegans* (GenBank accession CDH93266.1) with an E-value of 2.00E-04 (Table 1).

We next used the InterProScan software to search for conserved domains within the 18 GLAND proteins to gain further insight into their potential functions. Aside from N-terminal unintegrated signal peptides that were again predicted in all 18 GLAND proteins, InterPro domains were identified in only the four GLAND proteins 1, 13, 14, and 16 (Table 2). All domains found in these four GLAND proteins are consistent with their putative homologies (Table 1). Although InterPro domains were not identified in either GLAND15 or GLAND17, conserved domains from the Conserved Domains Database (CDD) were identified from our blastp searches using default parameters in both protein sequences. An MPN domain was identified in GLAND15 and an OTU-like cysteine protease domain in GLAND17 (Table 2).

**Table 2. Domains identified in six GLAND proteins**

| Clone | Descriptions of predicted domains (InterProScan or CDD)[a] |
|---|---|
| GLAND1 | GNAT (InterPro:IPR000182), Acyl-CoA N-acyltransferase (InterPro:IPR016181) |
| GLAND13 | Glycosyl hydrolase, family 32 (InterPro:IPR001362), Concanavalin A-like lectin/glucanases superfamily (InterPro:IPR008985), Glycosyl hydrolase family 32, N-terminal (InterPro:IPR013148), Glycosyl hydrolase family 32, C-terminal (InterPro:IPR013189), Glycosyl hydrolase, five-bladed beta-propellor (InterPro:IPR023296) |
| GLAND14 | Peptidase S9, prolyl oligopeptidase, catalytic domain (InterPro:IPR001375), Peptidase S9A, prolyl oligopeptidase (InterPro:IPR002470) |
| GLAND15 | MPN (CDD:cd08064) |
| GLAND16 | Chorismate mutase (InterPro:IPR002701), Chorismate mutase, type II (InterPro:IPR020822) |
| GLAND17 | OTU-like cysteine protease (CDD:pfam02338) |

[a] All eighteen GLAND proteins were analyzed with InterProScan for the presence of InterPro domains. GLAND proteins without InterPro domains were then analyzed with CD-search using blastp at NCBI to search for conserved domains from the Conserved Domains Database (CDD).

## 2.4 Discussion

A central question for plant-parasitic nematodes is how many of their estimated 15,000 to 20,000 protein-coding genes are directly involved in parasitism. Gao and colleagues (Gao et al. 2003) obtained a sizable profile of the *H. glycines* parasitome (i.e., the set of genes that encode the candidate effector proteins) that greatly added to the list of candidate effectors from any phytonematode that had been identified in prior studies (Ding et al. 1998; Smant et al. 1998; Lambert et al. 1999; Qin et al. 2000; Gao et al. 2001; Wang et al. 2001). Gao and colleagues identified fifty-one new candidate effectors at that time as being expressed specifically in the esophageal gland cells and that were all predicted to encode secreted proteins without TM domains, bringing the total *H. glycines* parasitome to sixty-four (Gao et al. 2003). This work resulted from the filtering of 2,229 high quality sequences that were derived from a gland cell-enriched cDNA library. Here, we generated and sequenced an *H. glycines* gland cell-enriched cDNA library that resulted in 3,392 high quality sequences (Fig. 1A), of which eighteen were determined to be new *H. glycines* candidate effectors based on the presence of N-terminal signal peptides for secretion, lack of TM domains and canonical ER localization signals, and gland-localized mRNA expressions. While this brings the total number of *H. glycines* candidate effector genes to around eighty-two, we have no way of confidently predicting the actual number of effector genes in the *H. glycines* parasitome. However, since we obtained over 1,000 more high quality sequences than Gao and colleagues (Gao et al. 2003), but accomplished a much lower discovery percentage for new effectors (0.53% vs.

2.29%), our study likely is very close to exhausting the discovery of additional *H. glycines* candidate effectors.

Our sequencing for this study was performed before the publication of our recent, more comprehensive approach where whole gland cells are isolated and subjected to next-generation sequencing (Maier et al. 2013), hence why here we still used the traditional approach for generating ESTs, consistent with Gao and colleagues (Gao et al. 2003). It will be interesting to compare the effectiveness of the method used by Gao and colleagues (Gao et al. 2003), and by us here, with that developed by Maier and colleagues (Maier et al. 2013) for the identification of new *H. glycines* effectors. Furthermore, it will be of particular interest to evaluate the relative abundances in the gland cells of previously identified candidate effectors, and the effectors reported here by us (*GLANDs 1-18*), as well as comparing these relative abundances between different *H. glycines* populations.

The paradigm for secretion of phytonematode effectors from the esophageal gland cells is the classical secretory pathway (Davis et al. 2004; Mitchum et al. 2013). In addition to aligning our 18 gland-positive cDNAs to the unpublished, draft *H. glycines* genome, we aligned the remaining 123 candidate cDNAs that did not result in gland-specific expression but that all had undergone and passed the same filtering process as the 18 GLAND genes, predicted the gene structure for each aligned scaffold and analyzed for the presence of N-terminal signal peptides. Interestingly, the majority of these non gland-specific cDNAs were not found to contain signal peptides for secretion. Since no cDNAs were found to be gland-specific in expression without simultaneously containing N-terminal signal peptides, we have found a perfect correlation between gland-specific

expression in the nematode and the presence of a predicted N-terminal signal peptide for secretion in our filtered sequence set. Therefore, these findings are in strong support for the classical secretory pathway being the predominant mode of secretion of phytonematode effectors from the gland cells.

Substantial portions of the parasitomes of phytonematodes have been demonstrated to consist of candidate effector sequences that lack detectable similarities to non-phytonematode, annotated sequences in databases, which are referred to as pioneers (Gao et al. 2003; Elling et al. 2009). It has been suggested that these pioneer candidate effectors are a result of faster-evolving sequences due to interactions with the host (Elling et al. 2009). Gao and colleagues (Gao et al. 2003) identified 51 *H. glycines* candidate effectors, of which 38 were designated as pioneers (75%). In our study, eleven of the eighteen new candidate effectors (61%) identified were either designated as pioneers, or showed similarity only to other phytonematode candidate effectors designated as pioneers. Only seven of the new candidate effectors identified resulted in similarity to non-pioneer sequences (Table 1) and/or contained predicted protein domains (Table 2). Thus, our findings reinforce the previous observations of pioneer candidate effectors predominating in phytonematode parasitomes. However, as our bioinformatic pipeline removed all sequences that significantly matched to proteins from *C. elegans*, it may be even more informative in future projects to test some of the sequences that would have otherwise survived the pipeline, for gland-cell mRNA accumulations.

Horizontal gene transfer (HGT) is well described for phytonematodes (Smant et al. 1998; Baldwin et al. 2004; Haegeman et al. 2011). It has been concluded that large suites of various non-effector and effector enzymes including those involved in cell wall-

degradation in phytonematodes were acquired from either plant-associated ancestral bacteria or fungi by HGT, since the phytonematode sequences had high similarity to enzymes from these types of organisms but no free-living or animal-parasitic nematodes were found to contain such enzymes (Baldwin et al. 2004; Haegeman et al. 2011). Here, we identified an *H. glycines* candidate effector (GLAND1) that is expressed in the dorsal gland during sedentary parasitic stages and that has high sequence similarity to GCN5-related N-acetyltransferases (GNATs) from bacterial species preferentially within the *Streptomyces* genus (Table 1). No such homologs are identified for the GLAND1 protein in other phytonematodes, animal-parasitic nematodes, or any free-living nematodes in blastp analysis against the NR database. However, more comprehensive analyses of phytonematode genomes must be performed in order to determine whether GLAND1 is present in only certain species, or more broadly conserved throughout phytonematodes. Because streptomycetes are saprophytes and, thus, in physical association with phytonematodes, it can be suggested that *GLAND1* might have been acquired via HGT from ancestral bacteria similar to streptomycetes. Furthermore, effector genes coding for acetyltransferases have been identified from a number of bacterial pathogens with functional roles in suppressing host immune responses (Lee et al. 2012; Paquette et al. 2012; Wu et al. 2012). However, effectors from the GNAT superfamily remain extremely rare. GNATs are the most widely distributed acetyltransferase systems functioning in diverse biological processes including antibiotic resistance (Draker and Wright 2004), sclerotization and neurotransmitter inactivation in insects (Han et al. 2012), various chromatin modifications (Jacobson and Pillus 1999), and regulation of polyamine metabolism (Bewley et al. 2006), among others. The GNAT effector from

*Mycobacterium tuberculosis* is the only pathogen effector belonging to the GNAT superfamily to be reported thus far (Kim et al. 2012). Similar to the *M. tuberculosis* GNAT effector, we speculate that GLAND1 may function by inhibiting defense signaling in plant cells during cyst nematode parasitism, but given the published roles of GNATs in a broad range of biological processes, as mentioned above, other functions are equally possible.

In plants, invertases (beta-fructofuranosidases) are crucial enzymes for metabolizing sucrose into glucose and fructose, as well as for sucrose transport (HaouazineTakvorian et al. 1997). Host plant invertases have been implicated in increasing the metabolic sink potential of giant-cells formed by the root-knot nematode *Meloidogyne incognita* (Kaplan et al. 2011), which is supported by the nematode's vital need for sucrose-derived carbohydrates such as fructose (Prasad et al. 2013). Also, invertases were identified in *M. incognita* during whole-genome sequencing and proposed to have been acquired from the plant symbiotic nitrogen-fixing bacteria of the genus *Rhizobium* through HGT due to highest similarity to sequences from these bacteria. It was hypothesized that these invertases serve to enhance the processing of plant-derived nutrients within the nematode, since no signal peptides for secretion were present in the protein sequences (Abad et al. 2008). On the other hand, recent whole-genome sequencing of *G. pallida* (potato cyst nematode) identified four genes that encode secreted invertases that are similar to the *Meloidogyne* invertases (Cotton et al. 2014). However, the possibilities of these secreted invertases in *G. pallida* being acquired through HGT from *Rhizobia* and being potential nematode effectors were not discussed nor was their gland-specific expression scrutinized. Here, we identified an *H. glycines*

secretory invertase (GLAND13), and like the predicted protein found in *M. incognita*, GLAND13 is most similar to invertase enzymes from bacteria, preferentially within the *Rhizobium* genus (Table 1). However, we were able to show that GLAND13 is expressed exclusively in the dorsal gland during the feeding stages of parasitism (Table 1), and thus, this *H. glycines* secreted invertase is likely an effector. This discovery may also suggest that the four secreted invertases in *G. pallida* are candidate effectors and likewise potentially acquired through HGT from *Rhizobium*. The possibility of these invertases being present in phytonematodes through HGT from *Rhizobium* is supported by the evidence that these bacteria were likely the predominant group of 'donor' bacteria partially responsible for nematode adaptations toward phytoparasitism (Scholl et al. 2003). However, since there has not yet been a deep phylogenetic analysis performed on phytonematode invertases, it still remains inconclusive whether or not the origins of these genes were from bacteria via HGT. Taken together, these results would suggest that cyst nematodes might increase the metabolic sink of syncytia by synthesizing and secreting their own plant nutrient-processing enzymes into the plant.

Animal-parasitic nematodes have been shown to secrete various proteases to break down host barriers during infection. For example, the entomopathogenic nematode *Steinernema carpocapsae* secretes a serine protease that facilitates invasion of the host insect gut wall, a physical barrier that opposes pathogen invasion (Toubarro et al. 2010). Although many proteases have been identified in phytonematodes, predominantly *M. incognita* (Abad et al. 2008), and have been shown to be expressed in the intestine (Neveu et al. 2003) and implicated in nematode reproduction and embryogenesis (Antonino de Souza Junior et al. 2013), phytonematode effectors secreted and delivered

into the plant that possess proteolytic activity have not been identified. Here, we identified two putative *H. glycines* candidate effector proteases, GLAND14 and 17, expressed specifically in the dorsal gland during later parasitic stages (Table 1). GLAND14 is most similar to an uncharacterized serine prolyl endopeptidase from the parasitic nematode *Ascaris suum* and is made up entirely of serine protease catalytic domains (Table 2). Since serine proteases are the most abundant proteolytic enzymes, are conserved throughout life, and possess diverse biological functions including digestion, fibrinolysis, development, blood coagulation, apoptosis, and immunity (Di Cera 2009), it is difficult to speculate the specific role that a serine protease effector would have during phytonematode parasitism, or whether host protein target(s) would be specific or more broad. However, since *GLAND14* resulted in mRNA accumulation in the dorsal gland cell in only late parasitic stages, we speculate that GLAND14 may target and cleave host plant proteins to weaken host defenses (Matas et al. 2014) rather than aid in the formation of the syncytium. Furthermore, sequence similarity and prediction of an OTU-like cysteine protease domain, which functions for deubiquitylation (Balakirev et al. 2003), in GLAND17 (Tables 1 and 2) might suggest that this candidate effector functions as a deubiquitylating enzyme during parasitism, which may be interesting given the reciprocal finding of ubiquitin-like effectors in phytonematodes (Gao et al. 2003; Guo et al. 2011; Wang et al. 2011; Chronis et al. 2013).

The MPN domain is diverse, found in subunits of multiprotein complexes from the 26S proteasome to the COP signalosome to subunits of eukaryotic translation intiation factor 3 (eIF3), as well as in regulators of transcription and translation (Sanches et al. 2007). The consensus of proposed biological functions is the involvement in protein

degradation, either through direct proteolytic activity, ubiquitin-binding for subsequent ubiquitylation, or direct association with the proteasome (Sanches et al. 2007). Therefore, since an MPN domain was identified in GLAND15, we propose participation in protein degradation within the syncytium as a tempting hypothesis for the function of this candidate effector.

Previously, an *H. schachtii* cellulose-binding protein (CBP) was shown to interact with a host pectin methylesterase (PME) for cooperative cell wall modification and ultimately for syncytium formation to promote parasitism (Hewezi et al. 2008a). Also, it was proposed that additional proteins might participate in the cell wall modifications enforced by the CBP-PME interaction (Hewezi et al. 2008a). Here, we identified a novel *H. glycines* candidate effector that is similar to phytonematode CBPs (Table 1). Although it is very distinct, GLAND10 has highest similarity with *H. schachtii* CBP, not *H. glycines* CBP. This finding of multiple, yet variable, CBP-like effectors in *H. glycines* demonstrates that the process of cooperative cell wall modification might involve multiple CBP effectors, which may or may not interact with different host proteins to carry out their functions. Furthermore, CBPs are proposed to have evolved from phytonematode cellulase or expansin cellulose-binding modules (Haegeman et al. 2011). Since both cellulases and expansins are believed to been acquired via HGT from bacteria, a phylogenetic analysis with the GLAND10 CBP, additional CBPs, cellulases and expansins from phytonematodes would be worthwhile for determining the origin of GLAND10 in the genome of *H. glycines*.

Much is known about the omnipresent pathogen effector chorismate mutase (CM). In plant parasitism, secreted CMs manipulate the host plant shikimate pathway.

CMs of root-knot and cyst nematodes (Bekal et al. 2003a, b; Doyle and Lambert 2003), as well as the maize smut fungus *Ustilago maydis* (Djamei et al. 2011) have been shown to alter plant cell development. It was proposed that these developmental alterations occurred likely through CM lowering auxin levels by causing a competition for chorismate, a central metabolite in the shikimate pathway (Bekal et al. 2003a; Doyle and Lambert 2003). Also, it was suggested that CM also suppresses the formation of plant defense compounds such as salicylic acid (Bekal et al. 2003a; Doyle and Lambert 2003). GLAND16 has highest similarity with the *H. glycines* CM, but is also much different in that its predicted CM domain (Table 2) makes up only a small portion of the protein. The majorities of the GLAND16 N- and C-termini are novel, whereas the previously described CM effector is mostly just a secreted CM domain (Bekal et al. 2003b). This discovery of a novel *H. glycines* candidate effector that has similarity to CMs of phytonematodes, but that is also highly different structurally, demonstrates that the process of manipulation of the host plant shikimate pathway during parasitism might be more complex than previously understood. Interestingly, the new *H. glycines* CM (GLAND16) putative protein sequence is 1,149-aa, which is over four times the size of the previously cloned phytonematode CMs and larger than all CMs cloned from any organism. Furthermore, because CMs in phytonematodes have also been proposed to been acquired via HGT from bacteria (Haegeman et al. 2011), a deep phylogenetic analysis for investigating the evolution of this novel CM (GLAND16) gene, in addition to GLAND1 and 13 discussed above, would be particularly interesting.

The syncytium is an elaborate feeding site established by cyst nematodes, most likely through the action of secreted products such as effector proteins (Hewezi and

Baum 2013; Mitchum et al. 2013). By exploring the molecular signaling between nematodes and their host plants, it is believed that it is possible to reveal vulnerable points during the parasitic cycle that can be exploited for the development of novel control measures (Hewezi and Baum 2013; Mitchum et al. 2013). For practical applications, interfering with these vulnerable points might allow the engineering of synthetic forms of resistance in the host plants to these nematodes. Furthermore, fundamental knowledge gained through *in planta* subcellular localization, interaction studies, and reverse genetic approaches for the candidate effectors identified here will greatly strengthen our understanding of how nematodes parasitize plants.

## 2.5 Authors' contributions

JBN is responsible for or significantly contributed to all analyses following the generation and initial bioinformatics analyses of the cDNA sequences. TRM provided extensive support to all nematological and molecular biology manipulations. TH contributed to bioinformatics analyses and guidance for molecular biology manipulations. ED, MGM, RH and TJB are responsible for the inception, development and guidance of this project and obtaining financial support. CS, JW, GW, VL and SD managed the data generation and bioinformatics analyses within DuPont Pioneer. TJB supervised the experimental work and co-wrote the manuscript with JBN.

# CHAPTER 3. HORIZONTAL GENE TRANSFER OF *ACETYLTRANSFERASES*, *INVERTASES* AND *CHORISMATE MUTASES* FROM DIFFERENT BACTERIA TO DIVERSE RECIPIENTS

A paper submitted to *Genome Biology and Evolution*

Jason B. Noon and Thomas J. Baum

## Abstract

Hoplolaimina plant-parasitic nematodes (PPN) are a lineage of animals with many documented cases of horizontal gene transfer (HGT). In a recent study, we reported on three likely HGT candidate genes in the soybean cyst nematode *Heterodera glycines*. These genes encode a GCN5-related N-acetyltransferase (GNAT; *Hg-GLAND1*), an invertase (INV; *Hg-GLAND13*), and a chorismate mutase (CM; *Hg-GLAND16*), all of which are most similar to bacterial sequences. Here, we tested possible HGT using deep Maximum Likelihood phylogenetic analyses combined with model selection tests of alternative tree topologies, which strongly supported HGT of all three genes from different rhizosphere bacteria to ancestral Hoplolaimina PPN and rejected descent via common ancestry. Mining of nematode databases determined that *GNATs* were transferred late in Hoplolaimina PPN evolution, while both *INVs* and *CM*s were transferred before Hoplolaimina divergence. While performing these analyses we surprisingly discovered additional HGT events of *GNATs*, *INVs* and *CMs* also from different bacterial donors to many other recipients. There were at least eleven and eight

well supported HGTs of *GNATs* and *INVs*, respectively, from different bacteria to diverse eukaryotes and archaea. Though less frequent, we also found data supporting the HGT of *CMs* from different bacteria to multiple eukaryotes. In other words, these three gene groups appear to have been frequent subjects of HGT from bacteria, which suggests that these genes may confer important evolutionary advantages to many taxa. In the case of Hoplolaimina PPN, this advantage likely was an improved ability to parasitize plants.

## 3.1 Introduction

Horizontal gene transfer (HGT) is common in bacteria and has recently been documented as an essential evolutionary process for many lineages of eukaryotes [reviewed in (Wijayawardena et al. 2013)]. In the phylum Nematoda (Fig. 1A), the plant-parasitic nematodes (PPN) of the suborder Hoplolaimina are among the eukaryotes with the most documented HGT events [reviewed in (Haegeman et al. 2011a)], especially for HGT from bacterial donors. For example, large suites of genes that encode plant cell wall-modifying proteins were determined to have been acquired in Hoplolaimina PPN via HGT from different bacterial donors (Danchin et al. 2010). Also, Hoplolaimina PPN were determined to have acquired enzymes for the vitamin B1, B5, B6 and B7 biosynthetic and salvage pathways, also from different bacterial donors (Craig et al. 2008; Craig et al. 2009). Furthermore, other genes in Hoplolaimina PPN are believed to have bacterial origins, but these hypotheses have not been rigorously tested [reviewed in (Haegeman et al. 2011a)].

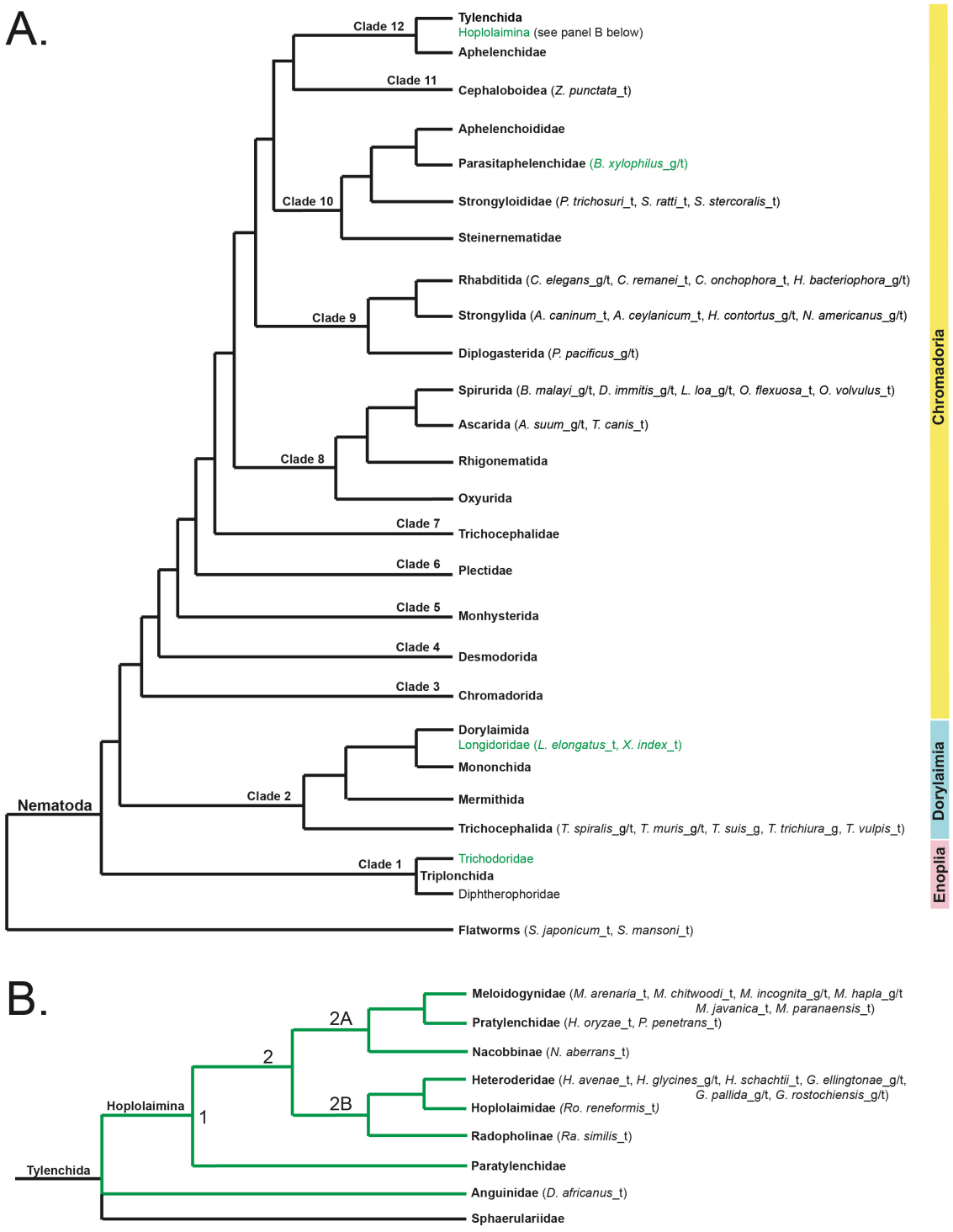



**Fig. 1. Cladograms of Nematoda and Hoplolaimina.** Tree topologies of the phylum Nematoda **(A)** and the suborder Hoplolaimina **(B)** are consistent with that described in Danchin et al. (2010) and are adapted from Holterman et al. (2006). **(A,B)** Nematode species whose genomic (_g), transcriptomic (_t), or both genomic and transcriptomic

(_g/t) sequences were included in our searches are listed in parentheses at each leaf. Branches that contain PPN species are illustrated in green.

In a recent study, we mined the secretory esophageal gland cells of *Heterodera glycines*, the soybean cyst nematode, to identify new candidate effectors (Noon et al. 2015). The esophageal gland cells have been shown to produce secretory proteins with signal peptides that are released into the nematode esophageal lumen and from there, delivered into plant tissues via a specialized hollow mouth spear, the stylet. Candidate nematode effectors are proteins produced specifically in these gland cells with N-terminal signal peptides for secretion into plant hosts [reviewed in (Mitchum et al. 2013)]. In that study, we identified candidate *H. glycines* effectors Hg-GLAND1 (GenBank: AJR19769.1), Hg-GLAND13 (GenBank: AJR19781.1) and Hg-GLAND16 (GenBank: AJR19784.1) whose predicted protein sequences exhibited significant similarities to proteins from different bacteria (Noon et al. 2015). All three *Hg-GLAND* genes were identified in an *H. glycines* draft genome and found to contain spliceosomal introns, which indicated that they were not prokaryotic contaminants.

A blastp search of Hg-GLAND1 detected highest sequence similarity to GCN5-related N-acetyltransferases (GNATs) from actinomycetes, predominantly streptomycetes, suggestive of HGT (Noon et al. 2015). Also, Hg-GLAND1 contained a predicted GNAT domain (InterPro: IPR000182), which suggested functionality (Noon et al. 2015). GNATs are the largest known enzyme superfamily functioning in diverse biological processes and are present in bacteria, archaea and eukaryotes (Vetting et al. 2005).

Blastp searches of Hg-GLAND13 detected highest sequence similarity to invertases (INVs) (Noon et al. 2015), enzymes that function in the hydrolysis of sucrose into glucose and fructose and/or sucrose transport. Abad et al. (2008) also identified two genes in the root-knot nematode *Meloidogyne incognita* genome assembly that encoded putative INVs. However, the *M. incognita* protein sequences did not contain N-terminal signal peptides, suggesting that these INVs are not effectors but function within the nematode. In blastp searches, the *M. incognita* INVs were found to be most similar to INVs from *Rhizobium* spp., thus suggesting possible HGT. A more recent study by Cotton et al. (2014) identified four putative *INV* genes in the potato cyst nematode *Globodera pallida* genome assembly with predicted N-terminal signal peptides in the encoded proteins. However, there was no determination of whether these genes are expressed in the esophageal gland cells, and thus, could be involved in the nematode's interaction with its host plant. In Noon et al. (2015) we established that the *H. glycines* GLAND13 INV is a candidate effector that is most similar to INV proteins from *Rhizobium*. The latter observation suggested that the *G. pallida* INVs mentioned above, which also have signal peptides, might likewise be candidate effectors (Noon et al. 2015). Importantly, although *INVs* are assumed to have been horizontally acquired in PPN from *Rhizobium* [reviewed in (Haegeman et al. 2011a)], to our knowledge this assumption is based entirely on blast searches and has not been subjected to phylogenetic analyses, nor have alternative hypotheses been tested (e.g., descent via common ancestry).

Blastp searches of Hg-GLAND16 revealed highest sequence similarity to chorismate mutases (CMs) (Noon et al. 2015). CMs are common in bacteria, plants, fungi and apicomplexan parasites, but rare in animals. This is due to the presence of the

shikimate pathway, for which CMs convert chorismate into prephenate in the former organisms, and its absence in animals (Bentley 1990). There are two structural types of CMs: type 1 or AroH class, which is characterized by a trimeric pseudo α/β-barrel structure (Chook et al. 1993), and type 2 or AroQ class characterized by a dimeric α-helical structure (Lee et al. 1995). Interestingly, nematodes do not contain the shikimate pathway, but PPN encode effectors that contain type 2 CM domains (Lambert et al. 1999; Gao et al. 2003; Jones et al. 2003; Huang et al. 2004; Huang et al. 2005; Noon et al. 2015). Type 1 CM domains have not been reported from nematodes. PPN CMs are largely believed to participate in the suppression of plant defenses (Lambert et al. 1999; Bekal et al. 2003a; Jones et al. 2003; Haegeman et al. 2011b; Yu et al. 2011), and to a lesser extent, to induce developmental changes in host plant roots (Bekal et al. 2003b; Doyle and Lambert 2003). Also, CMs from *Burkholderia* spp. have been reported most recently as the best match for CM sequences from PPN (Haegeman et al. 2011b). Consequently, it has been assumed that CM genes were horizontally acquired in PPN from bacteria (Lambert et al. 1999; Jones et al. 2003; Haegeman et al. 2011b) [reviewed in (Haegeman et al. 2011a)]. However, similar to the INVs mentioned above, these assumptions have been based largely on blast searches and sequence alignments alone, while alternative hypotheses have not been tested (e.g., descent via common ancestry in eukaryotes). Furthermore, while the Hg-GLAND16 candidate effector matched most highly to the previously reported *H. glycines* CMs in blastp analyses and contained a predicted type 2 CM domain (InterPro: IPR002701) (Noon et al. 2015), Hg-GLAND16 is over four times the size of previously reported *H. glycines* CMs. This observation

indicated that the repertoire of CMs in PPN, or at least in *H. glycines*, is more complex than previously appreciated.

Here, we used a combination of bioinformatic, phylogenetic and statistical analyses to rigorously test whether Hg-GLAND1, 13, and 16 were the subject of HGT. We determined how widespread individual gene sequences are within PPN to pinpoint when HGT most likely occurred during evolution. Since all Hg-GLAND1, 13 and 16 homologs that are identified in Hoplolaimina PPN contain the equivalent functional domains mentioned above, we simply refer to them throughout the paper as GNATs, INVs and CMs, respectively. Furthermore, we researched organisms outside of Nematoda to identify likely homologs with the goal to identify putative donor and additional recipient organisms of HGT events. Finally, we tested different evolutionary models to explain presence or absence of gene sequences in different taxa. These analyses confirmed that all three gene groups were acquired from bacteria whose descendants are currently found in the rhizosphere. Surprisingly, we also discovered that homologous ancestral bacterial sequences for two of these gene groups (*GNATs* and *INVs*) were likely the subject of very extensive HGT from highly different bacterial donors to many diverse recipient lineages of eukaryotes and archaea. A similar conclusion can also be drawn for *CMs*, but only to a smaller extent.

## 3.2 Materials and Methods

### 3.2.1 Searches of nematode sequence databases

In order to identify putative homologs in other nematodes for *GNATs*, *INVs* and *CMs*, the encoded protein sequences previously identified from *H. glycines* (Noon et al. 2015) were used as queries for blastn (Altschul et al. 1997) and tblastn searches against the transcript contigs, isotigs and genes, as well as to the reads grouped by library databases at Nematode.net (Wylie et al. 2004). We performed searches against all available nematode clades as well as to Hoplolaimina PPN separately. As outgroups in these searches, we included the genesets from flatworms and *Homo sapiens*. Also, we performed tblastn and blastp searches against all nucleotide and protein databases available at Nematodes.org, including NEMBASE4 (Elsworth et al. 2011), using an E-value threshold of 1E-04 (the online server did not allow 0.001). Furthermore, we performed tblastn searches against the raw sequence data obtained from the following published or unpublished transcriptome and genome assemblies: *Heterodera avenae* transcriptome (Kumar et al. 2014), *G. pallida* genome and transcriptome (Cotton et al. 2014), *Globodera rostochiensis* genome and transcriptome (Eves van-den Akker et al., unpublished), *Globodera ellingtonae* genome and transcriptome (Phillips et al., unpublished), *Nacobbus aberrans* transcriptome (Eves-van den Akker et al. 2014), *R. reniformis* transcriptome (Eves van-den Akker et al., unpublished), *Hirschmaniella oryzae* transcriptome (Eves van-den Akker et al., unpublished), and *Longidorus elongatus* transcriptome (Jones et al., unpublished).

### 3.2.2 Searches of NCBI sequences databases

To search for putative, non-nematode homologs of the GNATs, INVs, and CMs, the *H. glycines* homologs were used as queries for blastp (Altschul et al. 1997) searches against the following databases at the National Center for Biotechnology Information (NCBI): non-redundant protein sequences (nr), reference proteins (refseq_protein), patented protein sequences (pat), metagenomic proteins (env_nr), and transcriptome shotgun assembly proteins (tsa_nr). Separate blastp searches were performed specifically against the following taxids for each database: eukaryota taxid 2759, bacteria taxid 2, and archaea taxid 2157. We also searched the expressed sequence tag (EST) database at NCBI using the tblastn algorithm. All searches allowed for 1000 max target sequences and used an Expect (E-value; *E*) threshold of 0.001. Taxonomic classifications of the resulting similar sequences were evaluated using NCBI's taxonomy reports tool, which implements the taxonomy database at NCBI (Wheeler et al. 2008).

In order to maximize our sampling of the above databases for putative homologs of the protein families analyzed, we performed the same searches by using as queries the top bacterial protein sequences that matched most significantly to the respective *H. glycines* proteins. This search greatly increased the quantity of sequences and diversity of taxa that were included in our datasets for more comprehensive phylogenetic analyses.

### 3.2.3 Sequence retrieval

All sequences that aligned greater than 50 amino acids within the predicted protein domains of the *H. glycines* proteins (i.e., GNAT, GH32 INV and CM domains), as well as from the best-matching bacterial proteins, with E-values less than 0.001 were

kept and inspected for taxonomic classification. All nucleotide sequences obtained from transcriptomic, EST and EST contig databases were translated into protein sequences with the ExPASy translate tool. All genome assembly contigs from Hoplolaimina PPN were subjected to gene model and protein predictions using the self-training eukaryote gene prediction software GeneMark.hmm (Lomsadze et al. 2005) using the test set from the *C. elegans* genome. For non-nematode taxa, one to ten of the top scoring sequences from each taxonomic group were selected for multiple sequence alignments (MSAs), which allowed us to include a large quantity of sequences from organisms that were distantly related to Hoplolaimina PPN for phylogenetic analyses. No limit was set for the number of Hoplolaimina PPN sequences and all were included in the alignments.

### 3.2.4 Multiple sequence alignments

Sequence collections were uploaded into the sequence editor suite of the molecular evolutionary genetics analysis 6 (MEGA6) (Tamura et al. 2013) program. MSAs were performed using the program MUSCLE (Edgar 2004) with default parameters. Sequences that contained substantial gaps with poor alignments to otherwise high quality aligned regions were removed from the analysis in order to maximize the number of informative sites for phylogenetic analysis. Whenever a sequence was removed or edited from an original MSA, the MSA was systemically recalculated. The final MSAs were manually examined using the program Jalview (Waterhouse et al. 2009).

### 3.2.5 Phylogenetic analyses

We performed phylogenetic analyses using bootstrapped Maximum Likelihood (ML). To obtain the most reliable model of amino acid evolution we performed model selection analysis on MSAs using default parameters in the MEGA6 program (Tamura et al. 2013). For each protein family analyzed, the evolutionary model that resulted in the lowest Bayesian Information Criterion score was used (Tamura et al. 2013). Phylogenetic analyses were performed in MEGA6 using ML estimation with 100 bootstrap replications. Reported are the best-scoring ML phylogenetic trees with bootstrap values indicated on the corresponding nodes.

For each protein family analyzed, sequences that resulted in poorly supported clusters, contained relatively long branch lengths, and decreased the confidence of clusters overall within the respective phylogenetic trees were removed. Whenever sequences were removed, MSAs were recalculated, model selection analyses were repeated, and ML phylogenies were re-estimated accordingly. The resulting phylogenetic trees were initially annotated within MEGA6, and then detailed annotations were performed in Adobe Illustrator for visual purposes. The raw phylogenetic trees for each protein family are available in Figures S1-S5, Appendix B, and include identifications for all sequences used.

### 3.2.6 Model selection tests of alternative evolutionary hypotheses

For each phylogenetic tree presented in the paper, we generated alternative tree topologies from protein MSAs similar to (Theobald 2010) using the Topology Editor tool in MEGA6 (Tamura et al. 2013) in order to rigorously test alternative evolutionary

hypotheses. Taxa were placed into monophyletic groups according to their taxonomic classifications as reported in the taxonomy database at NCBI. Model selection analysis was performed using default parameters on both the original, unconstrained and constrained trees. Reported in Table 1 are the best scoring models of amino acid substitution, the number of parameters associated with the best model, and the Bayesian Information Criterion and corrected Akaike Information Criterion scores presented as the difference (Δ) from the unconstrained evolutionary hypotheses. Hypotheses that resulted in a difference in BIC and AICc scores of 5 or greater were considered as very strong empirical evidence for the better model (Burnham et al. 1998; Theobald 2010) [in this work lower scores are better (Tamura et al. 2013)]. Each model selection analysis was repeated at least once and we found that the results were identical in all trials.

### 3.2.7 %GC content and codon usage comparisons

We calculated %GC content for each coding DNA sequence (cds) using the formula ([(G+C)÷(G+C+A+T)]×100). Details for the number of cds included in each distribution, database sources and corresponding cds accession or identification numbers, %GC content for each cds, counts for each %GC content category, placement of confidence intervals, statistics of the distributions, and complete descriptions of how each distribution was generated are provided in Table S1. The %GC contents and accession or identification numbers for GNATs, INVs and CMs are provided in Table S2. The final distributions shown in Figure 7 were constructed in JMP Pro version 10.0.2 and were aligned for comparison purposes using Adobe Illustrator. We also analyzed and compared codon usages between the cds of the HGT candidates with codon usage tables

from both Hoplolaimina PPN and donor bacteria using the codon adaptation index (Puigbo et al. 2008). Complete details for the procedure, accession or identification numbers, codon usage tables used, observed and expected codon adaption indexes, and interpretation of the resulting values are provided in Table S3, Appendix B.

### 3.2.8 Searches for signal peptides, transmembrane regions and protein domains

To search simultaneously for secretion signal peptides and transmembrane (TM) regions, we used a combination of SignalP 4.1 (Petersen et al. 2011), TMHMM 2.0 (Krogh et al. 2001) and Phobius (Kall et al. 2007). For SignalP, we used the default D-cutoff values, but implemented both methods—SignalP-TM (input sequences may include TM regions) and SignalP-noTM (input sequences do not include TM regions). For TMHMM and Phobius, we used default parameters. To search for protein domains throughout our study, we used a combination of blastp and CD-search (Marchler-Bauer and Bryant 2004) at NCBI to search the conserved domains database (CDD) (Marchler-Bauer et al. 2015), and InterProScan 5 (Jones et al. 2014) to search the InterPro protein families database (Mitchell et al. 2015). The InterProScan searches also allowed another round of predictions for signal peptides and TM regions.

## 3.3 Results and Discussion

### 3.3.1 GNAT, INV and CM homologs are found only in nematodes within the Hoplolaimina suborder

The phylum Nematoda is composed of 12 major clades (Fig. 1A) (Holterman et al. 2006). Hoplolaimina is found in clade 12 within the order Tylenchida, which in addition to Hoplolaimina contains the basal plant-pathogenic Anguinidae and entomopathogenic Sphaerulariidae nematodes (Fig. 1B). Also in clade 12 is the fungal-feeding family Aphelenchidae, and immediately basal to clade 12 is the clade 11 superfamily Cephaloboidea containing strictly bacterial-feeders (Fig. 1A). Furthermore, Hoplolaimina contains strictly plant parasites, and this suborder is subdivided into the larger clade 1 containing all Hoplolaimina PPN, and the smaller clade 2 with family Paratylenchidae as outgroup (Fig. 1B). Hoplolaimina clade 2 is further subdivided into clades 2A and 2B (Fig. 1B). Hoplolaimina clade 2A contains root-knot (family Meloidogynidae; *Meloidogyne* spp.), lesion (family Pratylenchidae) and false root-knot (family Nacobbinae) nematodes (Fig. 1B). Hoplolaimina clade 2B contains cyst (family Heteroderidae), reniform (family Hoplolaimidae) and burrowing (family Radopholinae) nematodes (Fig. 1B).

As a first step in our analyses, we performed a comprehensive search of available nematode genome and transcriptome assemblies to identify homologs of the three candidate HGT genes in question (i.e., *GNATs*, *INVs* and *CMs*) in parasitic and non-parasitic nematode species other than *H. glycines* (Fig. 1). Our searches included extensive genome and/or transcriptome assembly sequences from Nematoda clades 2 and

8-12, and all of the Hoplolaimina clade 2 families mentioned above, totaling 46 different nematode species. One nematode species included only genome assembly sequences, 26 included only transcriptome assembly sequences, and 18 included both genome and transcriptome assembly sequences. For those nematode species that only included transcriptome assembly sequences, we cannot rule out the possibility that lack of gene identification is due to lack of gene expression, rather than gene absence entirely. Noteworthy, multiple species within Nematoda clades 2, 8 and 10, as well as multiple species within Hoplolaimina clades 2A and 2B, included both genome and transcriptome assembly sequences (Fig. 1A). In these analyses, all three candidate HGT genes were identified to different degrees in Hoplolaimina PPN (Fig. 2), as further described below, but we did not find any significant nematode matches ($E < 0.001$) outside of this suborder. This included the lack of identification within Nematoda clades 2 and 8-11 (Fig. 1), as well as the Anguinidae lineage immediately basal to Hoplolaimina (Figs. 1B and 2); however, the latter only included transcriptome assembly sequences. Thus, these results indicated the scarcity of the three candidate HGT genes in question throughout Nematoda, and suggested that these genes might only be present within Hoplolaimina.

Within Hoplolaimina, *GNATs* were completely absent from the three PPN families within clade 2A, while a single *GNAT* homolog was found in cyst and reniform nematodes within clade 2B (Fig. 2A). Different from *GNATs*, multiple *INV* (Fig. 2B) and *CM* (Fig. 2C) homologs were identified throughout both Hoplolaimina clades 2A and 2B PPN. Thus, these findings indicated that the *GNATs* appeared in Hoplolaimina clade 2B after the divergence from Hoplolaimina clade 2A, while *INVs* and *CMs* appeared before the divergence of Hoplolaimina clade 2.

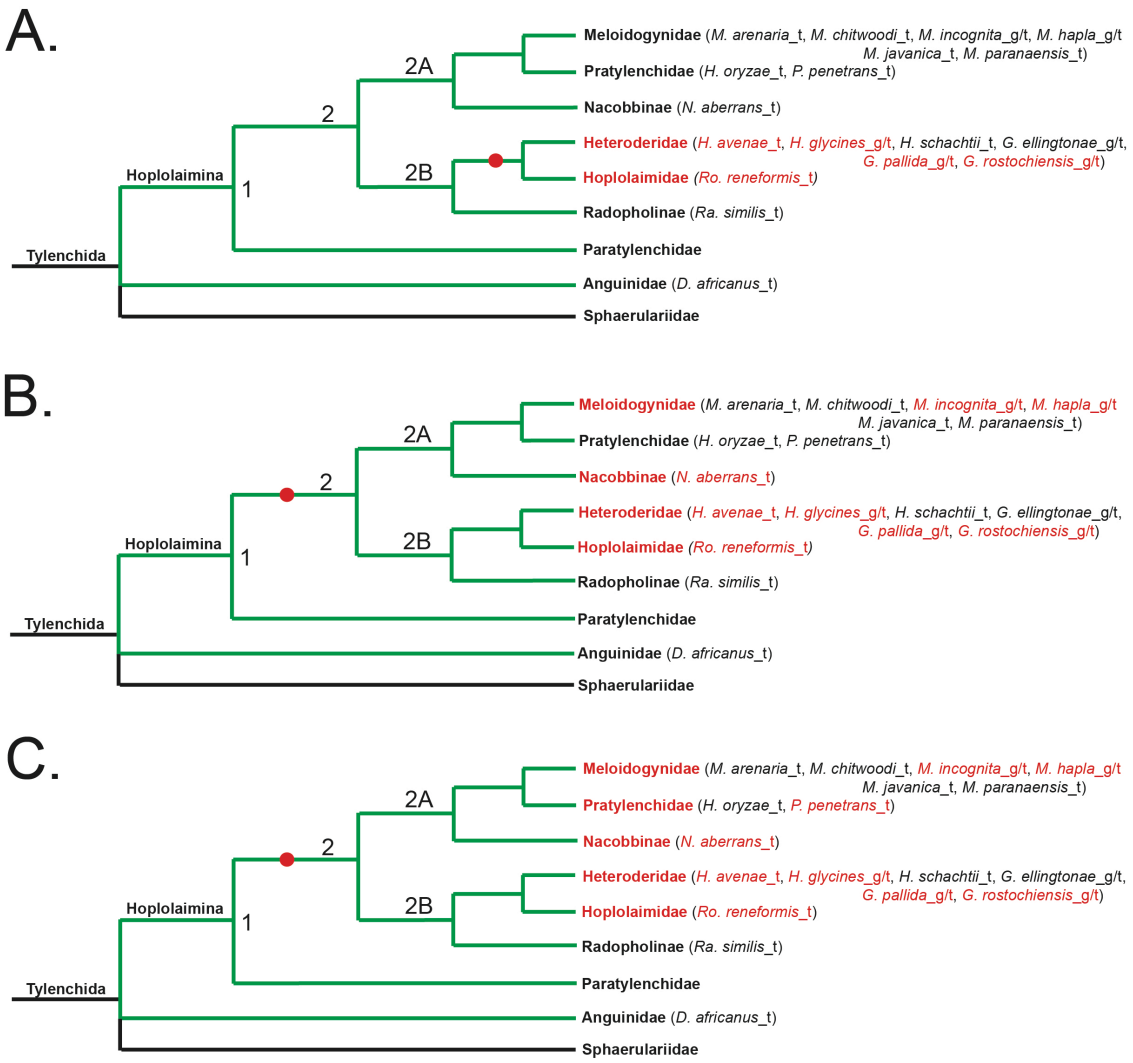



**Fig. 2. Timing of appearance of *GNATs*, *INVs* and *CMs* in Hoplolaimina PPN. (A-C)** Cladograms are shown as in Figure 1B. Species that were found to contain homologs of the HGT genes in question are colored red. The suspected timing of appearance of *GNATs* **(A)**, *INVs* **(B)** and *CMs* **(C)** are illustrated with a red circle placed on the appropriate branch.

Although it was evident that the *GNATs* appeared in Hoplolaimina clade 2B PPN after divergence from Hoplolaimina clade 2A, *Radopholus similis* is the only species with sequences to represent the burrowing nematodes, and only has limited transcriptome

assembly sequences (Fig. 2A). Thus, it remains possible that *GNATs* could also be present within this Hoplolaimina clade 2B basal lineage.

Due to insufficient representation of genome and transcriptome assemblies basal to Hoplolaimina, it was not possible to predict the precise appearance of *INVs* and *CMs* within the distal Nematoda clades (Fig. 1). Though it was evident that these two candidate HGT genes appeared before the divergence of Hoplolaimina clade 2, we did not have sequences for the basal Paratylenchidae lineage within Hoplolaimina (Fig. 2B,C). Also, we only had limited transcriptome assembly sequences for the Anguinidae lineage basal to Hoplolaimina, and no sequences were available for Sphaerulariidae or the other clade 12 Aphelenchidae lineage (Fig. 1). Furthermore, transcriptome assembly sequences were only available for a single species within the clade 11 Cephaloboidea lineage (Fig. 1A). On the other hand, sufficient transcriptome assembly sequences were included for 4 nematode species representing 2 different lineages within clade 10, and the facultative plant-parasitic species *Bursaphelenchus xylophilus* included both genome and transcriptome assembly sequences, thus providing rigorous support for the absence of *INVs* and *CMs* from this clade and the more basal Nematoda clades. Thus, it remains possible that *INVs* and *CMs* could be present throughout clades 11 and 12 nematodes, and better sequence representation for these lineages in the future will determine the precise conservation of these candidate HGT genes.

### 3.3.2 Hoplolaimina GNATs, INVs and CMs cluster with bacteria in phylogenetic analyses

We had determined that the three candidate HGT genes in question (i.e., *GNATs*, *INVs* and *CMs*) were specific to members in the Hoplolaimina PPN lineage and likely absent from all other nematodes. Next, to test the hypothesis that all three candidate HGT genes were horizontally acquired in Hoplolaimina PPN, we performed blastp searches to identify all possible homologs in every NCBI protein sequence database as well as the EST database in order to conduct the most comprehensive phylogenetic analyses possible. If the Hoplolaimina sequences were to cluster with similar bacterial sequences over other similar eukaryotic, or even archaeal sequences, this would support HGT over vertical inheritance. Finally, we used model selection analyses to compare the likelihoods of HGT versus descent via common ancestry in order to provide the most rigorous support for one evolutionary scenario over the other.

For our blastp searches, we used the complete Hoplolaimina GNAT, INV and CM protein sequences as queries, and the protein sequence hits with similarities of $E < 0.001$ to the predicted functional domains were considered as potential homologs, and were thus used in downstream phylogenetic analyses. The majority of protein sequence hits from these analyses were from bacteria. Thus, in order to maximize our sampling of protein sequences from eukaryotes and archaea, in addition to bacteria, we performed separate blastp searches using the bacterial homologs as queries and also considered the resulting non-bacterial protein sequence hits as potential homologs for downstream phylogenetic analyses. For both Hoplolaimina GNATs and INVs, we identified hundreds of potentially homologous sequences covering all three domains of life (eukaryotes,

archaea and bacteria), while for CMs, potentially homologous sequences were only found in bacteria and a few other eukaryotes.

We made a particularly interesting discovery when all GNAT sequences discovered by these searches were analyzed. GNATs have been reported to fall into one of the following six families based on sequence, structure and function (although no extensive phylogenetic analyses have been reported to date): bacterial aminoglycoside N-acetyltransferases (NATs), animal serotonin NATs, actinobacterial mycothiol synthases, bacterial Fem aminoacyltransferases, eukaryote glucosamine-6-phosphate NATs, and eukaryote histone acetyltransferases [reviewed in (Vetting et al. 2005)]. Thus, before testing the HGT hypothesis, we were interested in determining which GNAT family the Hoplolaimina GNATs belong to. Blastp searches using Hoplolaimina GNATs or their most similar bacterial sequences (i.e., actinomycete GNAT sequences) as queries revealed significant similarities ($E < 0.001$) to protein sequences from other bacteria and archaea, as well as to other eukaryotes. Because no phylogenetic analyses had been reported for GNATs to date, we constructed a ML phylogenetic tree that included the Hoplolaimina GNATs and their blastp hits identified by us along with a large number of known representatives from all six GNAT families. As expected, this analysis showed that all six known GNAT families formed highly supported monophyletic groups (Fig. 3). However, this analysis also resulted in a seventh, highly supported monophyletic group for all Hoplolaimina GNAT sequences along with all bacterial, archaea and other eukaryotic GNAT sequences identified in our blastp searches (Fig. 3, Novel GNAT Family). These findings strongly suggested that Hoplolaimina GNATs and their blastp matches form a novel, seventh GNAT family that has not been described. It can also be

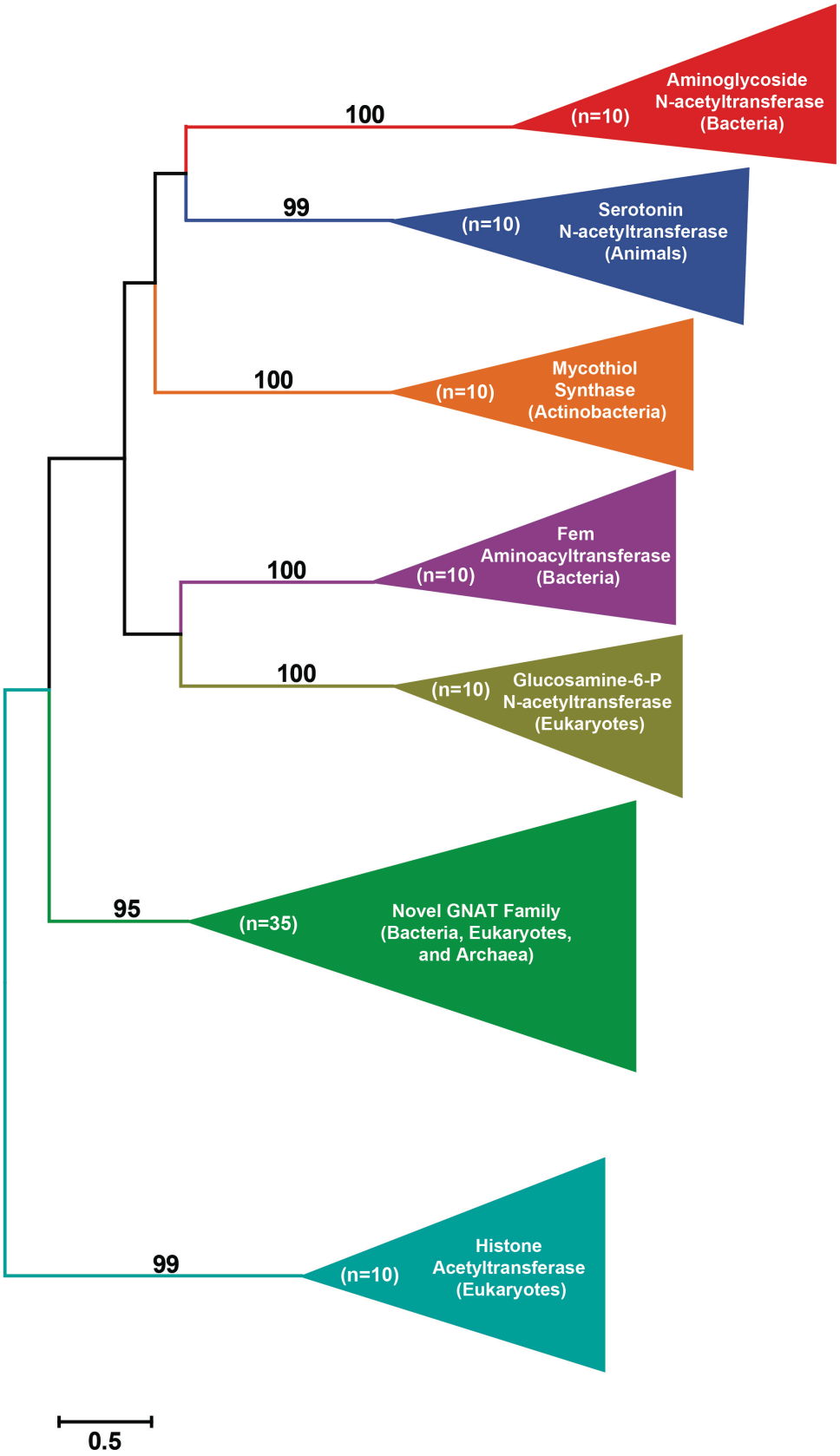

100
(n=10)
Aminoglycoside
N-acetyltransferase
(Bacteria)
99
(n=10)
Serotonin
N-acetyltransferase
(Animals)
100
(n=10)
Mycothiol
Synthase
(Actinobacteria)
100
(n=10)
Fem
Aminoacyltransferase
(Bacteria)
100
(n=10)
Glucosamine-6-P
N-acetyltransferase
(Eukaryotes)
95
(n=35)
Novel GNAT Family
(Bacteria, Eukaryotes,
and Archaea)
99
(n=10)
Histone
Acetyltransferase
(Eukaryotes)
0.5

**Fig. 3. Phylogenetic tree of the GNAT superfamily and newly identified GNATs similar to the Hoplolaimina homologs.** Phylogenetic groups containing each GNAT family are collapsed and color-coded with corresponding bootstrap support values indicated at each node. The number of sequences (n) that were used for each GNAT family is indicated within each collapsed phylogenetic group. Organisms that contain each GNAT family are provided in parentheses within each collapsed phylogenetic group. Note that the newly identified GNAT clade with similarity to the Hoplolaimina homologs forms a highly supported monophyletic group with no significant clustering to any other GNAT family, thus indicating a novel GNAT family, which we called Family 7 (FAM7). The raw phylogenetic tree is shown in Figure S1, Appendix B.

speculated that the lack of clustering of this seventh GNAT family to the other six GNAT families suggests that these sequences are not GNATs. However, prediction of functional GNAT domains in all sequences of the seventh cluster, including all Hoplolaimina GNATs, suggests otherwise, and thus, we refer to the collection of these sequences throughout the rest of the paper as Family 7 (FAM7) GNATs.

In the ML phylogenetic tree of FAM7 GNATs (Fig. 4), which contained over one hundred sequences, Hoplolaimina clustered with actinomycetes (we included streptomycete sequences since these are the bacterial sequences that are most similar to the Hoplolaimina sequences). Although the bootstrap support for the cluster containing streptomycetes and Hoplolaimina (labeled cyst nematodes in Figure 4) is not highly supported (bootstrap = 58), the next closest node supporting the larger cluster of cyst nematodes, streptomycetes, the actinomycete *Tetrasphaera japonica* and leotiomycete fungi is well supported (bootstrap = 81). Also, within this cluster *T. japonica* and leotiomycete fungi are in a highly supported cluster (bootstrap = 90), lending additional support for the cluster containing streptomycetes and Hoplolaimina (Fig. 4).

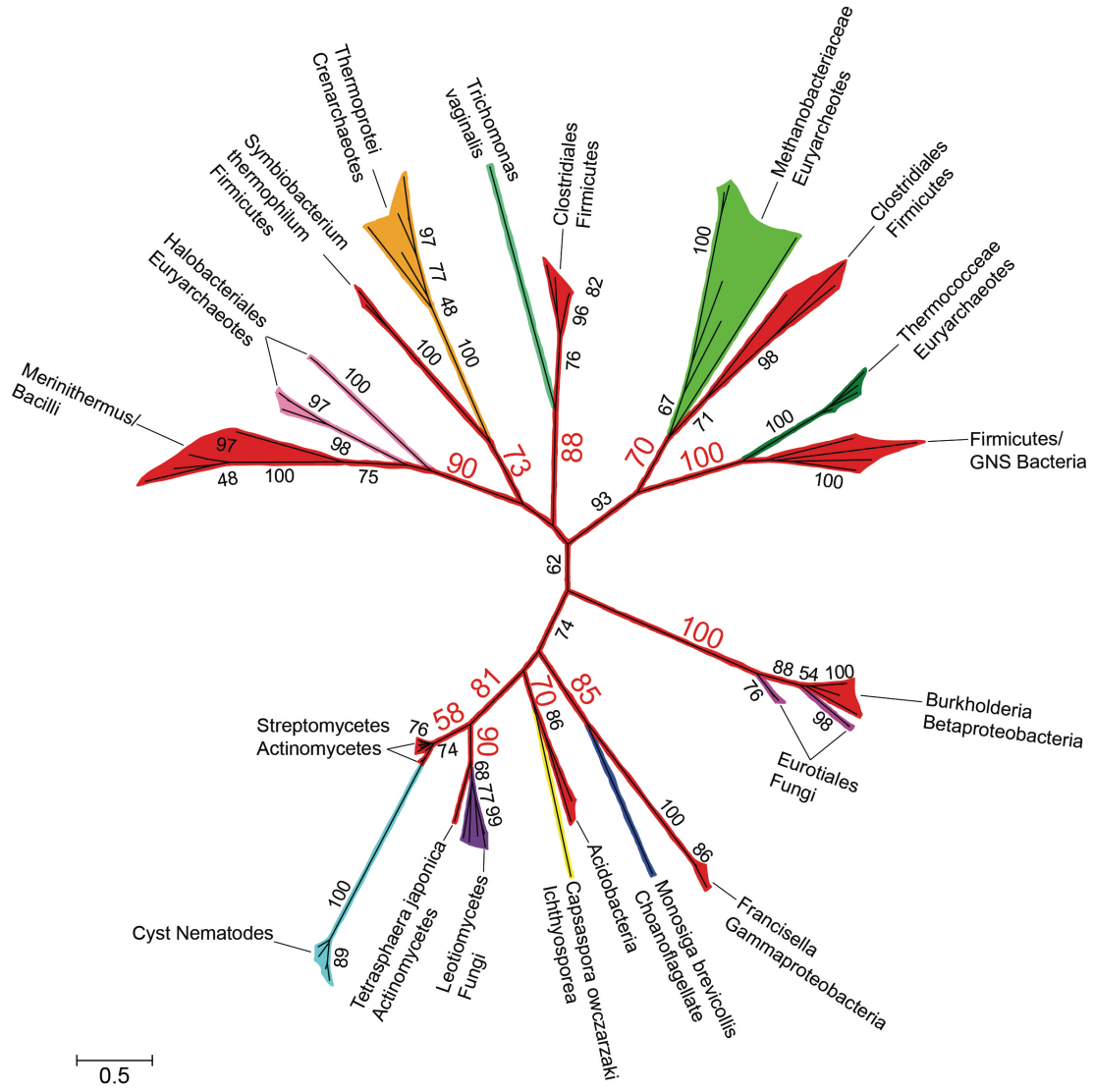



**Fig. 4. Phylogenetic tree of FAM7 GNATs including the Hoplolaimina homologs.** Phylogenetic groups are color-coded according to their taxonomic classifications. Bootstrap support values are indicated at corresponding nodes, and those that support possible HGT events are oversized in red font. Notice a maximum of 10 possible HGT events where eukaryotes and archaea form monophyletic groups with different bacteria, including cyst nematodes with actinomycetes most similar to streptomycetes. The raw phylogenetic tree is shown in Figure S2, Appendix B.

Similarly, Hoplolaimina clustered with rhizobacteria (order Rhizobiales) with very strong support (bootstrap = 100) in the ML phylogenetic tree of INVs (Fig. 5), which also contained over one hundred sequences.

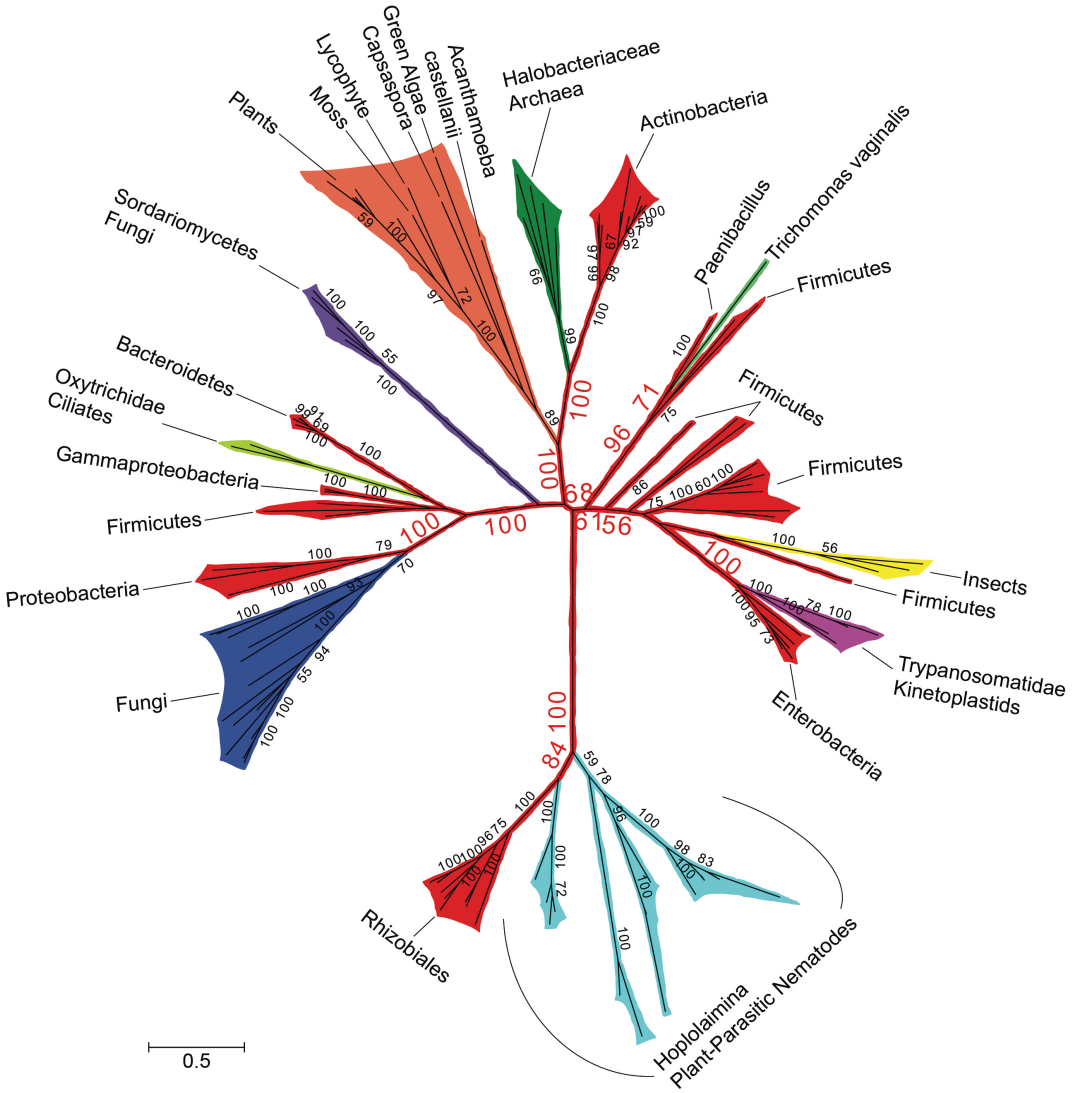



**Fig. 5. Phylogenetic tree of INVs similar to the Hoplolaimina homologs.** Phylogenetic groups are color-coded according to their taxonomic classifications. Bootstrap support values are indicated at corresponding nodes, and those that support possible HGT events are oversized in red font. Notice a maximum of 8 possible HGT events where eukaryotes and archaea form monophyletic groups with different bacteria, including Hoplolaimina PPN with rhizobacteria (order Rhizobiales). The raw phylogenetic tree is shown in Figure S3, Appendix B.

Finally, in the ML phylogenetic tree of CMs (Fig. 6), which was much smaller than the FAM7 GNAT and INV phylogenetic trees, but still included all possible homologs that were identified in other eukaryotes from the NCBI sequence databases, PPN CMs formed a supported cluster (bootstrap = 77) with *Burkholderia* CMs.

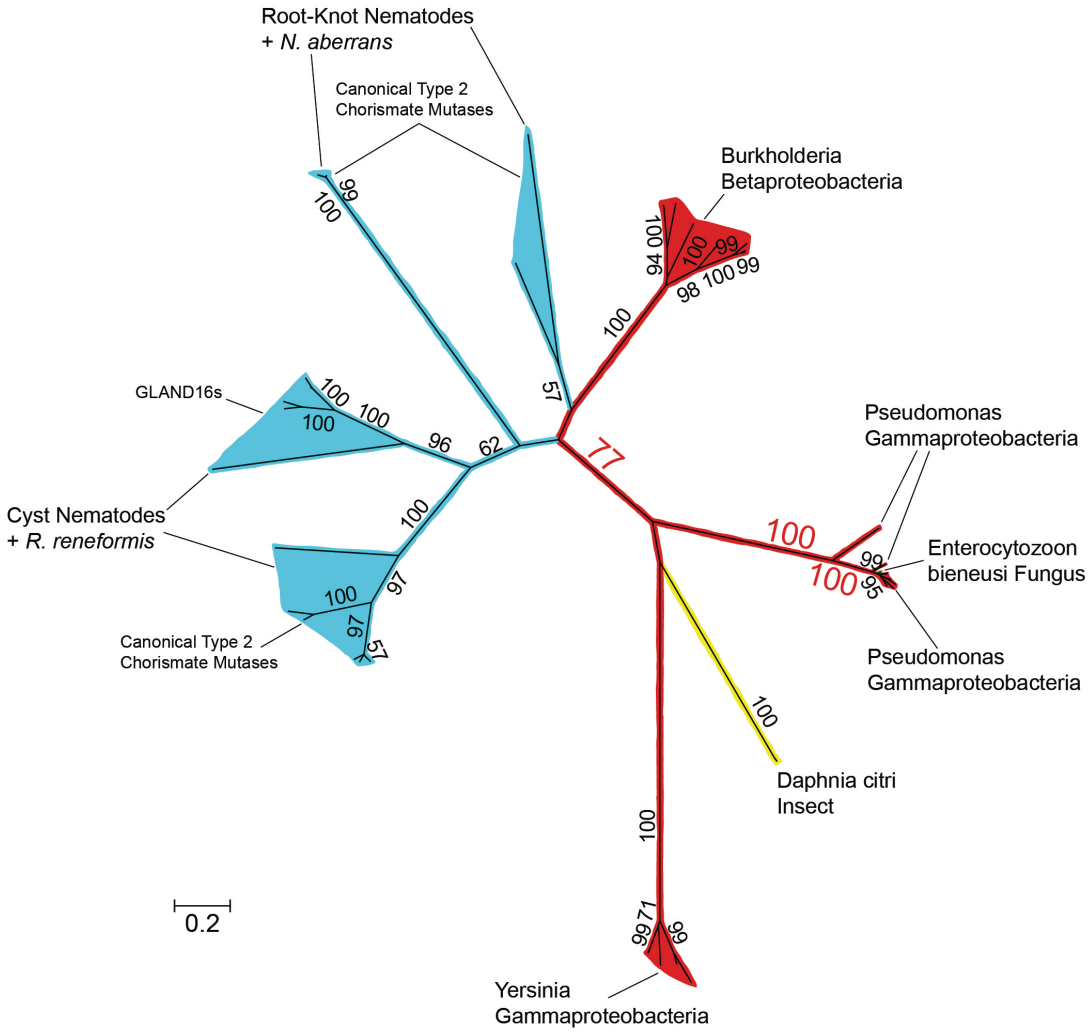



**Fig. 6. Phylogenetic tree of CMs similar to the Hoplolaimina homologs.** Phylogenetic groups are color-coded according to their taxonomic classifications. Bootstrap support values are indicated at corresponding nodes, and those that support possible HGT events are oversized in red font. Notice a supported monophyletic grouping of Hoplolaimina PPN with *Burkholderia* CMs. The raw phylogenetic tree is shown in Figure S4, Appendix B.

Taken together, these results strongly supported the hypothesis that all three candidate HGT genes were horizontally acquired in Hoplolaimina PPN from bacteria. Importantly, all three suspected bacterial donors are commonly found in the rhizosphere, and thus in the same niche as Hoplolaimina PPN. The latter findings document a physical association between the putative donor and recipient organisms, which further supported our HGT hypothesis.

### 3.3.3 *FAM7 GNATs*, *INVs* and *CMs* were horizontally acquired in Hoplolaimina from rhizosphere bacteria

The analyses described above determined that the three groups of nematode effector proteins in question cluster with protein sequences of the suspected donor bacteria in phylogenetic analyses that included all possible homologs that can be found in NCBI protein sequence and EST databases. However, phylogenetic analyses alone are insufficient to document HGT, as descent via common ancestry cannot be completely ruled out using this method. Model selection analysis is a formal method for comparing the likelihoods of different models of evolution such as HGT versus descent via common ancestry and has been used to test the hypothesis of a universal common ancestry of life (Theobald 2010). In model selection analysis, hypothesized trees, constrained by chosen criteria, are constructed for a given sequence alignment, and models of amino acid substitution and the associated scores [in our case, Bayesian and corrected Akaike Information Criteria (BIC and AICc, respectively)] are calculated. This analysis therefore provides a rigorous method for testing HGT versus descent via common ancestry, and thus, we employed this methodology here to test HGT of the candidate Hoplolaimina

genes. For all three candidate HGT genes, the unconstrained HGT models consisted of the trees that resulted from our phylogenetic analyses (Figs. 4-6). For constrained models that were consistent with descent via common ancestry, Hoplolaimina PPN were grouped with taxa according to known taxonomic classifications (Table 1). For each unconstrained and constrained model of evolution, the rank of score, constraint used (if any), model of amino acid substitution that resulted from the analysis, number of parameters used in each analysis, and the resulting BIC and AICc scores expressed as the difference from the unconstrained model are reported in Table 1.

For each of the three candidate HGT genes in question, the unconstrained and all constrained models of evolution resulted in very similar models of amino acid substitution and number of parameters (Table 1). Since the BIC and AICc scores for each model are weighted by both the likelihood and number of parameters used, the differences in scores observed for each model of evolution represent almost exclusively differences in likelihoods rather than differences in the complexities of each model (Theobald 2010). Accordingly, the unconstrained HGT models for all three candidate HGT genes scored substantially lower (lower scores are better) than all constrained models of evolution that were consistent with descent via common ancestry (Table 1). Models with even the subtlest constraints placed on the unconstrained HGT models resulted in substantially higher scores. For example, placing the Hoplolaimina FAM7 GNATs with leotiomycete fungi rather than with streptomycete FAM7 GNATs (see Fig. 4), and placing Hoplolaimina CMs with insect rather than with *Burkholderia* CMs (see Fig. 6), resulted in substantially higher scores compared to the unconstrained HGT models (Table 1). These results indicated that the rhizosphere bacteria with which the

candidate HGT genes clustered in the phylogenetic analyses (Figs. 4-6) actually were the HGT sources of the *FAM7 GNAT*, *INV* and *CM* genes in Hoplolaimina.

**Table 1. Model selection tests of constrained versus unconstrained models of evolution for the candidate Hoplolaimina HGT genes**

| Evo. Model | Rank | Constraint | Sub. Model | K | ΔBIC | ΔAICc |
|---|---|---|---|---|---|---|
| | | *FAM7 GNATs* | | | | |
| unconstrained | 1 | . | LG+G | 154 | 0 | 0 |
| constrained 1 | 8 | Euk+Arch | LG+G | 154 | 1175.733 | 1175.697 |
| constrained 2 | 6 | Euk | LG+G+I | 155 | 634.763 | 627.339 |
| constrained 3 | 7 | CN+Fungi+Mon+Cap | LG+G | 154 | 648.002 | 647.967 |
| constrained 4 | 4 | CN+Fungi+Mon | LG+G | 154 | 549.181 | 549.146 |
| constrained 5 | 5 | CN+Fungi | LG+G | 154 | 580.907 | 580.872 |
| constrained 6 | 2 | CN+L Fungi | LG+G | 154 | 126.561 | 126.525 |
| constrained 7 | 3 | CN+E Fungi | LG+G | 154 | 229.615 | 229.580 |
| | | *INVs* | | | | |
| unconstrained | 1 | . | WAG+G | 232 | 0 | 0 |
| constrained 1 | 7 | Euk+Arch | WAG+G+I | 233 | 2149.761 | 2140.827 |
| constrained 2 | 6 | Euk | WAG+G+I | 233 | 1894.372 | 1885.438 |
| constrained 3 | 5 | PPN+Insects+ Fungi+P/M/L/C/G/A+ Excavates | WAG+G+I | 233 | 1665.100 | 1656.167 |
| constrained 4 | 4 | PPN+Insects+Fungi+ P/M/L/C/G | WAG+G+I | 233 | 945.151 | 935.950 |
| constrained 5 | 3 | PPN+Insects+ Fungi | WAG+G+I | 233 | 822.972 | 814.039 |
| constrained 6 | 2 | PPN+Insects | WAG+G+I | 233 | 229.779 | 220.545 |
| | | *CMs* | | | | |
| unconstrained | 1 | . | WAG+G+I | 89 | 0 | 0 |
| constrained 1 | 3 | Euk | WAG+G+I | 89 | 129.280 | 129.281 |
| constrained 2 | 2 | PPN+Insects | WAG+G+I | 89 | 26.109 | 26.110 |

K=number of parameters, G=Gamma distributed rate variation among amino acid positions, I=invariant amino acid positions, Euk=eukaryotes, Arch=archaea, CN=cyst nematodes, PPN=Hoplolaimina plant-parasitic nematodes, Mon=*Monosiga brevicollis*, Cap=*Capsaspora owczarzaki*, P/M/L/C/G/A=Plants/Mosses/Lycophytes/*Capsaspora owczarzaki*/Green Algae/*Acanthamoeba castellani*.

### 3.3.4 Evolution of *FAM7 GNATs*, *INVs* and *CMs* in Hoplolaimina following HGT from rhizosphere bacteria

After determining that the three nematode effector genes in question were horizontally acquired in Hoplolaimina PPN from the respective rhizosphere bacteria, we tested whether these genes resembled %GC contents and codon usages similar to the donor or to the recipient genomes. For %GC content, we collected cds for all recipient Hoplolaimina PPN and donor bacteria (Table S1) in order to generate distributions of %GC content for each (Fig. 7). %GC contents were calculated for members of each of the three Hoplolaimina HGT gene families (Table S2) followed by an evaluation for placement of the calculated %GC contents on each distribution (Fig. 7). Nearly all members evaluated from each of the three HGT gene families resulted in %GC contents similar ($P > 0.05$) to the recipient Hoplolaimina genomes and significantly different ($P < 0.05$) from the donor bacterial genomes (Fig. 7). Only two Hoplolaimina *INVs*, one from *G. pallida* and the other from *N. aberrans*, resulted in %GC contents significantly different ($P < 0.05$) from recipient Hoplolaimina and similar ($P > 0.05$) to donor bacterial genomes (Fig. 7B, *Gp* and *Na*).

For codon usage analyses, we calculated codon adaptation indexes (CAIs) and compared them with the expected CAIs (E-CAIs) (Puigbo et al. 2008). Similar to %GC content, we found that nearly all members of the three HGT gene families used codons that were significantly similar ($P < 0.05$) to Hoplolaimina genomes and different from donor bacterial genomes (Table S3, Appendix B). Taken together, these results indicated that subsequent to HGT, all three acquired gene families experienced adaptation to the recipient Hoplolaimina genomes. These findings were consistent with the current

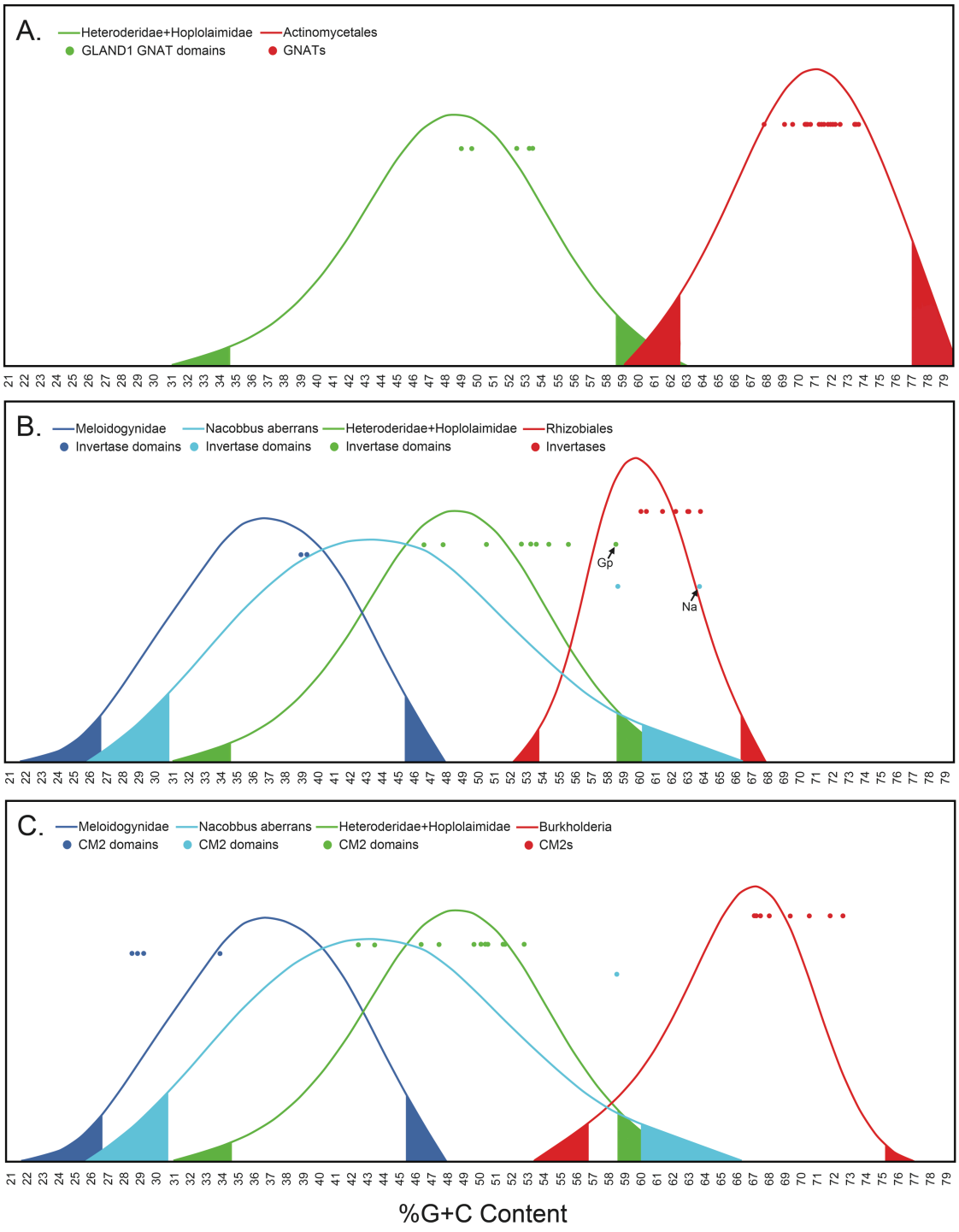



**Fig. 7. %GC content comparisons of Hoplolaimina HGT genes with distributions constructed from recipients and donors.** Distributions of %GC content were constructed using cds from each respective group of Hoplolaimina and donor bacteria listed in each panel. The height of each distribution corresponds to the number of cds at that particular value of %GC content. The x-axis is labeled at the bottom with %GC

content. Dots toward the top of each distribution indicate the %GC content for the functional domain (transferred form) of each HGT gene. Dots are included for the donor bacterial genes as reference. Tails on each distribution correspond to the upper and lower limits of two-tailed 95% confidence intervals. All raw data are provided in Tables S1 and S2.

paradigm for HGT in PPN that in order for transferred genes to be functional in recipient genomes, they must adapt for efficient transcription and translation (Scholl et al. 2003; Danchin et al. 2010).

From our above searches for the three HGT genes in nematode genomes and transcriptomes, we found potentially complex patterns of gene losses in Hoplolaimina following HGT, especially for the *INVs*. However, we considered that some of these alleged gene losses could have been due to insufficient sequence databases, as already mentioned. Also, as mentioned above, INVs are currently understood to be non-secreted in root-knot nematodes, while in cyst nematodes they are believed to be secreted effectors. Furthermore, previously reported CMs are relatively small proteins that have been documented in all Hoplolaimna PPN with considerable sequence datasets. However, GLAND16 CMs are over four times larger than other CMs (Noon et al. 2015) and in our above searches were only found in cyst nematodes. Therefore, we were interested in elucidating the complex post-HGT evolution of these gene families in Hoplolaimina. In these analyses, we evaluated the subtrees of the Hoplolaimina recipients and bacterial donors specifically within the ML phylogenetic trees that resulted from our comprehensive phylogenetic analyses, re-evaluated the multiple sequence alignments, and evaluated the protein sequences for predicted functional domains, signal peptides and TM regions. Results from these analyses are detailed in the supplementary text, Appendix B. In summary, results from these analyses indicated that a *FAM7 GNAT* was acquired

from actinomycetes in an ancestor of cyst and reniform nematodes and remains as a single effector gene in each species. Also, *INVs* and *CMs* were acquired from rhizobacteria and *Burkholderia*-related bacteria, respectively, in ancestral Hoplolaimina PPN, and since HGT have experienced multiple duplications with subfunctionalization; most are likely effectors while minorities are, interestingly, likely TM proteins functioning within the nematodes.

### 3.3.5 Bacteria were likely HGT hubs of *FAM7 GNATs*, *INVs* and *CMs* to diverse recipients

As mentioned above, from our blastp searches of NCBI protein sequence and EST databases for possible non-nematode homologs of the Hoplolaimina FAM7 GNATs and INVs, we identified numerous possible homologs from all three domains of life (bacteria, archaea and eukaryotes). A total of sixteen different eukayote or archaea lineages were found to contain possible homologs of Hoplolaimina FAM7 GNATs, all of which clustered together in the phylogenetic tree of the GNAT superfamily (Fig. 3). For the INVs, we found a total of nine different eukaryote or archaea lineages that contained possible homologs of the Hoplolaimina INVs. Although to a much smaller extent, we found two additional eukaryotes with possible homologs to Hoplolaimina CMs. Interestingly, all suspected eukaryote and/or archaea homologs of the HGT genes in question matched to different lineages of bacteria in the blastp searches, and most formed well-supported clusters with the different bacteria in the phylogenetic trees (Figs. 4-6; Fig. S5, Appendix B). There were a total of eleven such examples of different eukaryote and archaea lineages forming well-supported clusters with different bacteria for the

FAM7 GNATs (Fig. 4; Fig. S5, Appendix B), eight such examples for INVs (Fig. 5), and three such examples for CMs (Fig. 6). Moreover, in the NCBI sequence databases, all three HGT genes were found to be present in essentially all bacteria, but only in relatively few eukaryotes and archaea with no indication of common ancestors containing the genes. These findings contradict vertical inheritance being responsible for the presence of any of these genes in these diverse lineages of eukaryotes and archaea. Rather, these findings are best explained by multiple independent HGT events from numerous different bacteria to diverse recipients.

Also important was the finding that many of the suspected donor bacteria (or at least their descendants in cases of more ancient HGTs) for all three HGT genes occupy niches that are very similar to those of the recipient organisms, similar to what we described above for soil dwelling bacterial donors and Hoplolaimina PPN. For example, leotiomycete fungi are commonly found in the soil, and like Hoplolaimina PPN, their suspected *FAM7 GNAT* donors are actinomycete soil bacteria. *Trichomonas vaginalis* and the Clostridiales firmicute *FAM7 GNAT* donor—we mostly identified *Lachnospira multipara*—are found in the human urogenital tract and human intestine, respectively. Also, Trypanosomatids are often found in insects, and the best matches of the Trypanosomatid protein sequences were to the Enterobacteria *Providencia* spp., bacteria that are found in the haeomolymph of some insects. Furthermore, the following bacterial donors to archaea are even more consistent with occupying similar niches: Firstly, Halobacteriales euryarchaeotes consist of extreme halophiles, and one of the most similar bacteria was *Alkalibacillus haloalkaliphilus*, also an extreme halophile. Secondly, Methanobacteriaceae euryarchaeotes consist of extremophiles (‘methanophiles’), and the

most similar bacterium was *Dethiobacter alkaliphilus*, also an extremophile (probably halophile). Thirdly, Thermococceae euryarcheatoes are extreme thermophiles, and the most similar bacteria were *Coprothermobacter proteolyticus* and *Thermotoga hypogea*, both extreme thermophiles. Lastly, Thermoprotei crenarchaeotes are also extreme thermophiles, and among the most similar bacteria were *Thermobaculum terrenum* and *Symbiobacterium thermophilum*, again also extreme thermophiles. Taken together, these findings indicated that the majority of all donors and recipients of the HGT genes in question occupy similar niches, which further strengthens the conclusion of numerous, independent HGTs.

## 3.4 Conclusions

In this study, using a combination of deep phylogenetic analyses and tests of alternative evolutionary hypotheses, we have determined that three gene families in Hoplolaimina PPN were acquired via HGT from different rhizosphere bacteria. These three gene families are the *GLAND1s* (which encode proteins that were determined to be part of a novel family of GNATs which we called FAM7), *INVs* and *CMs*. Some of the homologs from each HGT gene family have evolved into *bona fide* or candidate effectors subsequent to HGT. A *FAM7 GNAT* was acquired in the Hoplolaimina clade 2B lineage from actinomycetes most similar to streptomycetes and presently encodes the GLAND1 candidate effector in cyst and reniform nematodes. Similarly, *INV* and *CM* genes were acquired in Hoplolaimina from rhizobacteria and *Burkholderia*-related bacteria, respectively, but before the divergence of clade 2. Subsequent to HGT, the acquired *INV*

and *CM* genes appear to have experienced complex duplications with subfunctionalization (e.g., some homologs presently encode candidate or *bona fide* effectors, and some encode TM proteins likely functioning within the nematodes).

Remarkably, we also found that *FAM7 GNATs*, *INVs*, and to lesser extent *CMs*, were likely subjects of numerous HGTs from bacteria to diverse recipients, including both eukaryotes and archaea for the former two genes. The suspected donors for nearly all HGTs occupy very similar niches as the recipient organisms, thus strengthening the conclusion of numerous possible HGTs. These findings indicate that bacteria likely served as hubs for HGT of these three genes to diverse recipients, and demonstrate their likely importance for not just Hoplolaimina PPN, but for many diverse taxa.

## 3.5 Authors' contributions

JBN designed and performed all analyses, and wrote the manuscript. TJB supervised the work and co-wrote the manuscript with JBN.

# CHAPTER 4. GENOMIC ASSOCIATION OF *CER* RETROVIRUSES WITH HORIZONTALLY ACQUIRED GENES IN CLADE 12 PLANT-PARASITIC NEMATODES

A paper submitted to *Genome Biology and Evolution*

Jason B. Noon and Thomas J. Baum

## Abstract

There are many examples of horizontal gene transfer (HGT) to plant-parasitic nematodes (PPN) within clade 12 Nematoda. However, how foreign DNA was transferred into the genomes of these PPN, and then vertically transmitted, is completely unknown. In some eukaryotes, HGT genes are overrepresented in chromosomal regions enriched with transposable elements (TEs), thus leading to hypotheses that TEs were involved in HGT, although association between TEs and horizontally transferred genes is not evident in clade 12 PPN. Most TEs in nematodes are *BEL/Pao* or *Ty3/Gypsy* LTR retrotransposons, collectively called *Cer* retrotransposons in the model nematode *Caenorhabditis elegans*. Intriguingly, a *Cer* retrotransposon was previously shown to have acquired an *envelope* gene from Phleboviruses (single-stranded ambisense RNA viruses), resulting in the formation of *Cer* retroviruses capable of infection and intercellular movement in *C. elegans*. This discovery is of particular relevance when considering that Phleboviruses have been shown to infect clade 12 PPN, raising the possibility that *Cer* retroviruses are also present in these nematodes. Such infectious

particles can be hypothesized to have mediated the capture and delivery of foreign DNA, intercellular movement (including to the germ line), and integration into the recipient genome. In support of this hypothesis, we document the presence of *Cer* retroviruses integrated into the genomes of the cyst nematodes *Heterodera glycines* and *Globodera pallida*. Moreover, we observed many documented HGT genes in close proximity to genome-integrated *Cer* retroviruses. Searches of over one hundred randomly selected genomic scaffolds from both nematode genomes determined that the association of HGT genes with *Cer* retroviruses is highly significant. The genomic association discovered here provides a starting point for further investigation into the mechanisms of HGT in PPN.

## 4.1 Introduction

Horizontal gene transfer (HGT) is common in bacteria, and examples of eukaryotes having acquired a complement of their genes horizontally are ever increasing. Most documented HGTs in eukaryotes involved transfers from microbial donors. Examples include small-scale HGTs such as the transfer of only two *invertase* (*beta-fructofuransidase*) genes from bacteria to the emerald ash borer *Agrilus planipennis* (Zhao et al. 2014), to more large-scale HGTs such as massive transfers from bacteria and fungi to Lepidopteran insects (Sun et al. 2013) and bdelloid rotifers (Gladyshev et al. 2008). Also, though more rare, HGT has occurred in the opposite direction (i.e., from higher eukaryote donors to microbial recipients). For example, bacteria and fungi acquired *expansin* genes from plants that increased the ability of these microbes to loosen

plant cell walls for better colonization (Nikolaidis et al. 2014). Thus, it is clear that HGT is an essential evolutionary process for many eukaryotes (Wijayawardena et al. 2013), and numerous, additional examples will likely be documented.

Holterman et al. (2006) divided the phylum Nematoda into 12 major clades (Fig. 1). The clade 12 plant-parasitic nematodes (PPN; order Tylenchida; suborder Hoplolaimina) are among the eukaryotes with the most documented HGTs (Haegeman et al. 2011). A large suite of genes that encode plant cell wall-modifying proteins was acquired in clade 12 PPN via HGT from different bacterial donors (Danchin et al. 2010). Also, these PPN acquired many different genes that encode enzymes for the vitamin B biosynthetic and salvage pathways, also from different bacterial donors (Craig et al. 2008; Craig et al. 2009). Interestingly, multiple genes were transferred from *Rhizobium* to root-knot nematodes (*Meloidogyne* spp.) (Scholl et al. 2003), and a *polyglutamate synthase* gene was transferred from bacteria to *Meloidogyne artiellia* (Veronico et al. 2001). The latter findings possibly indicate that HGT occurred so recent to be present within a single genus of clade 12 PPN, or maybe even a single species. Furthermore, we determined recently that genes encoding GCN5-related N-acetyltransferases, invertases and chorismate mutases were horizontally acquired from rhizosphere bacteria (Noon and Baum, manuscript submitted). A number of other genes in clade 12 PPN are believed to have microbial origins (Haegeman et al. 2011), but mostly, these hypotheses have not been rigorously tested.

Despite the increasing prevalence of HGT in eukaryotes, the mechanistic details remain elusive. Even if a horizontal acquisition of microbial genes by eukaryotic cells becomes further plausible, the mechanisms by which such genes arrive in the germ line to

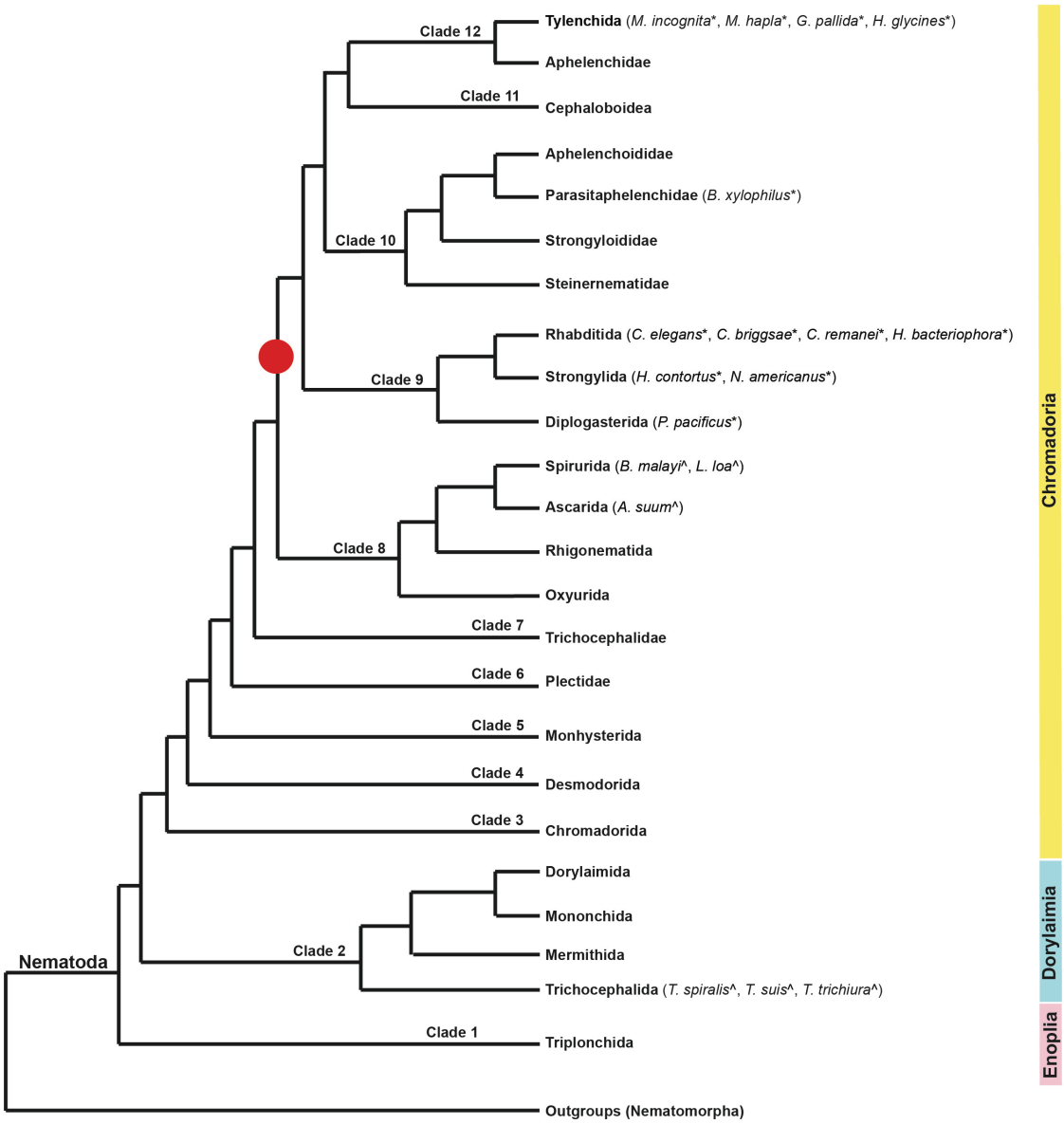



**Fig. 1. Cladogram of the phylum Nematoda.** Clades are consistent with Holterman et al. (2006). The three nematode subclasses, Enoplia, Dorylaimia and Chromadoria are indicated to the right of the corresponding clades. Nematode species whose genome sequences were included in our searches for *Cer* retroviruses are indicated in parentheses at the corresponding leaves. *, *Cer* retrovirus found; ^, *Cer* retrovirus not found. The red circle indicates when the *Cer* retroviruses most likely appeared in the phylum Nematoda—in the common ancestor of clades 9 through 12.

be vertically transmitted are completely unknown. Clade 12 PPN are obligate biotrophs of mostly roots that live in close proximity to many of the suspected bacterial donors in

the rhizosphere (Haegeman et al. 2011; Noon and Baum, manuscript submitted). This proximity should be a minimum requirement for HGT to occur. However, our current understanding is that these PPN do not feed on bacteria at any stage of their life cycles and are not infected by them, so it is still difficult to envision how genetic material was internalized and how the actual HGT would have occurred. Interestingly, in the phylogeny of Nematoda, bacterial-feeding Cephaloboidea nematodes (clade 11) are placed immediately basal to clade 12, which contains Tylenchida and the fungal/oomycete-feeding yet plant-parasitic Aphelenchidae (Fig. 1) (Holterman et al. 2006; Blaxter and Koutsovoulos 2015; Quist et al. 2015). Consequently, the current theory is that HGT of cell-wall modifying protein genes, and possibly other HGT genes, occurred in ancestral bacterial-feeders that were similar to the Cephaloboidea nematodes (Quist et al. 2015). Acquisition of these genes probably allowed nematodes to degrade fungal/oomycete cell walls for consumption, and as a result, led to the emergence of the fungal/oomycete-feeders that were similar to Aphelenchidae nematodes (Quist et al. 2015). Finally, PPN emerged within clade 12 from the ancestral fungal/oomycete-feeders (Fig. 1). Therefore, the ingestion of 'donor' bacteria by ancestral bacterial-feeding nematodes would have provided direct contact of bacterial genetic material with nematode digestive systems, thus enabling HGT by so far unknown mechanisms. However, several documented HGT genes were transferred after the diversification of clade 12 PPN (Veronico et al. 2001; Noon and Baum, manuscript submitted), and thus, the latter theory would not at all explain how these more recent HGTs occurred. Moreover, even the more ancient HGTs in the ancestral bacterial-feeding nematodes would still not explain how the foreign DNA moved from the intestine into the germ line

and then integrated into the recipient genome. Therefore, it is clear that *(i)* the mechanism underlying HGT in clade 12 PPN remains largely obscure, and *(ii)* that clade 12 PPN themselves must be equipped with a mechanism for HGT.

### 4.2 *Cer* retroviruses contain the necessary elements for HGT in nematodes

Although mechanisms for HGT in eukaryotes are unknown, transposable elements (TEs) like DNA transposons and retrotransposons are hypothesized to have been involved in some instances (Wijayawardena et al. 2013). For example, Gladyshev et al. (2008) found numerous HGT genes in bdelloid rotifers that were clustered at chromosomal regions near the telomeres where TEs are enriched. Thus, genomic association of HGT genes in bdelloid rotifers with TEs led to the hypothesis that TEs contributed to HGT at least in these eukaryotes (Gladyshev et al. 2008).

What makes retrotransposons, in particular long-terminal repeat (LTR) retrotransposons, so compelling as possible mechanistic agents for HGT in eukaryotes is their integration mechanism (Wijayawardena et al. 2013) and potential for sequence homology-independent recombination (Varmus and Brown 1989; Malik et al. 2000). These mobile genetic elements are first transcribed and then reverse transcribed to form a complementary DNA, and the double-stranded DNA is ultimately integrated into a new genomic location via the encoded integrase. Also, during conversion to a double-stranded DNA, LTR retrotransposons, as well as retroviruses, undergo intermolecular strand-transfer events at their LTRs to both RNA and DNA templates, a mechanism that opens the door for homology-independent recombination (Varmus and Brown 1989; Malik et

al. 2000). Thus, it seems plausible that if foreign DNA or RNA were to enter a cell that harbors an active LTR retrotransposon, then that LTR retrotransposon could by chance form and illegitimate recombination intermediate with the foreign DNA or RNA before integrating back into the genome. Interestingly, this is how many retroviruses are formed—an LTR retrotransposon recombines with a viral *envelope* gene in the same cell, presumably by chance, before integrating back into the genome (Varmus and Brown 1989; Malik et al. 2000; Ganko et al. 2001). Addition of the viral *envelope* gene, which encodes surface glycoproteins, adds the infectious intercellular movement function to the LTR retrotransposon, thus earning its classification as a retrovirus (Malik et al. 2000; Ganko et al. 2001).

Although retrotransposons (and DNA transposons) are relatively scarce in nematode genomes, especially compared to humans and plants (Wright and Finnegan 2001), they are still present. Most retrotransposons in nematodes are from the *Ty3/Gypsy* or *BEL/Pao* LTR retrotransposon families (de la Chaux and Wagner 2011). In the free-living, model nematode *Caenorhabditis elegans* (clade 9; Fig. 1) *Ty3/Gypsy* and *BEL/Pao* retrotransposons are grouped into a larger family called the *C. elegans retroelements* or *Cer* elements (Malik et al. 2000; Ganko et al. 2001). *Cer* retrotransposons from the *BEL/Pao* family contain *gag* and *pol* genes combined into a single open reading frame (ORF) (Fig. 2A). The ORF is flanked by a tRNA primer-binding site, a polypurine tract and at the extreme ends, the 5'- and 3'-LTRs, which allow transcription (Malik et al. 2000). Within the ORF, *gag* encodes the nucleocapsid that contains a zinc finger CCHC-type RNA-binding domain (Fig. 2A). The *pol* gene encodes

the polyprotein that contains the reverse transcriptase, peptidase, and an RNase H/integrase region (Fig. 2A).

Interestingly, Dennis et al. (2012) detected Viral-Like Particles (VLPs) from a *BEL/Pao* family *Cer* retrotransposon called *Cer1* at high levels in the pachytene/diplotene region of the *C. elegans* hermaphrodite gonad (i.e., in the germ line). In the same study, *Cer1* was shown to translocate into germ cell nuclei. Also, Dennis et al. (2012) found VLPs at high levels in the gonads of *Caenorhabditis japonica*, but the identities of the VLPs were not determined. These data convincingly show the high activity levels of some *Cer* elements and most importantly, show their activity in the nematode germ line. As a result, such *Cer* elements can be discussed as being involved in mediating HGT.

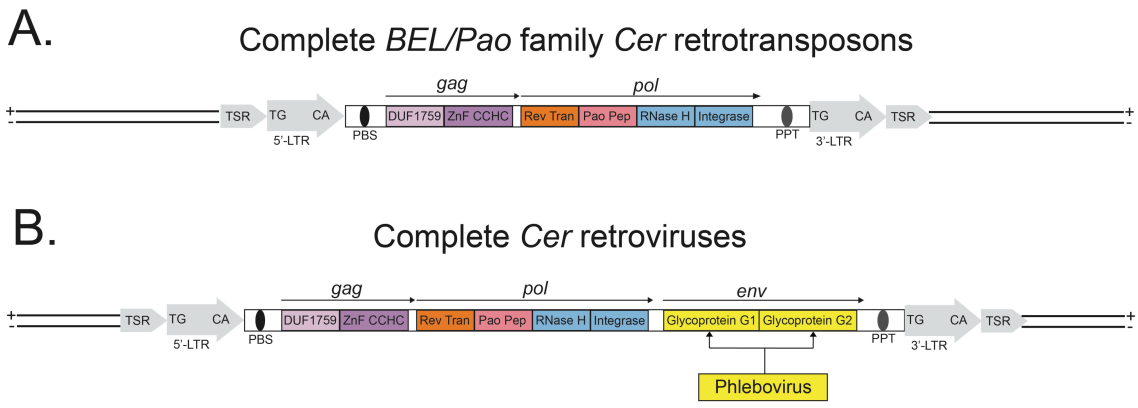



**Fig. 2. Schematic representation of the *BEL/Pao* family of *Cer* retrotransposons and *Cer* retroviruses. (A)** Structure of the *BEL/Pao* family of *Cer* retrotransposons (e.g., *Cer1*). **(B)** Structure of *Cer* retroviruses. **(A,B)** TSR, target site repeat; PBS, tRNA primer-binding site; DUF1759, domain of unknown function 1759; Rev Tran, reverse transcriptase; Pao Pep, BEL/Pao peptidase A17; PPT, polypurine tract.

However, Danchin et al. (2010) showed that HGT genes in clade 12 PPN, in particular those encoding the cell-wall modifying proteins, are not significantly associated with TEs, including the *BEL/Pao* family of *Cer* retrotransposons. Similar to the reports by Gladyshev et al. (2008) that TEs are associated with HGT genes in bdelloid rotifers, we would expect that if *BEL/Pao* family *Cer* retrotransposons such as *Cer1* contributed to HGT in clade 12 PPN, there should be a physical genome association of *Cer* elements and HGT genes. Therefore, it appears unlikely that *BEL/Pao* family *Cer* retrotransposons such as *Cer1* themselves actually contributed to HGT in clade 12 PPN.

But interestingly, during nematode evolution a *BEL/Pao* family *Cer* retrotransposon near identical to *Cer1* acquired an *envelope* gene from a Phlebovirus that resulted in the *Cer7*, *13* and *14* retroviruses in *C. elegans* (Malik et al. 2000 and Ganko et al. 2001). These three *Cer* retroviruses are almost identical in sequence, and thus, we simply refer to them throughout this paper as *Cer* retroviruses. The only difference between *BEL/Pao* family *Cer* retrotransposons and retroviruses is that the latter contain an *envelope* gene from a Phlebovirus at the 3'-end of the ORF (Fig. 2B). The *Cer* retrovirus *envelope* gene encodes the surface glycoproteins G1 and G2 from Phleboviruses (Fig. 2B), which permit infectious intercellular movement (Malik et al. 2000; Ganko et al. 2001). Phleboviruses are single-stranded ambisense RNA viruses within the family Bunyaviridae. Remarkably, Phleboviruses actually have been reported to infect clade 12 PPN (Bekal et al. 2011), meaning that these nematodes contain the extracellular receptors that Phlebovirus envelope glycoproteins bind to that allows for infection. This discovery raises the specter that *Cer* retroviruses are also present in clade 12 PPN, and thus, that *Cer* retroviruses may have formed in a common ancestor of

Nematoda clades 9 through 12 (see Fig. 1). This would likewise suggest that Phleboviruses may have infected these agronomically and medically relevant nematodes for quite some time and played an important role in their evolution. Also, *Cer* retroviruses might be capable of spreading throughout the nematode body, ultimately reaching the germ line and translocating into germ cell nuclei, like *Cer1* in *C. elegans*. Furthermore, it is then possible to envision *Cer* retroviruses capturing foreign DNA via intermolecular strand-transfer, invading the nematode germ line through cell-to-cell movement, and integrating the foreign DNA into the genome. Therefore, we hypothesized that *Cer* retroviruses might have been the mechanistic agents for HGT in clade 12 PPN, and possibly other nematodes.

To investigate this hypothesis, we first searched for *Cer* retroviruses in nematode genome sequences in order to estimate when *Cer* retroviruses might have appeared in Nematoda. Then, we inspected the DNA sequences nearby documented HGT gene families in clade 12 PPN genomes to observe for potential association with *Cer* retroviruses. As controls, the same searches were performed on over one hundred randomly selected genomic scaffolds.

### 4.3 *Cer* retroviruses are present in genomes of clades 9, 10 and 12 nematodes

While *Cer* retrotransposons are mostly conserved throughout Nematoda (de la Chaux and Wagner 2011), to the best of our knowledge, *Cer* retroviruses had only been reported in *C. elegans* (Malik et al. 2000; Ganko et al. 2001). Thus, we were first interested in revealing when *Cer* retroviruses appeared in Nematoda, which would also

determine when the Phlebovirus *envelope* gene was recruited to a *Cer* retrotransposon (Malik et al. 2000; Ganko et al. 2001). For this analysis, we tblastn-searched published and unpublished genome sequences from eighteen different nematode species that covered clades 2, 8, 9, 10 and 12 within Nematoda (Fig. 1). As query, we used Phlebovirus glycoprotein G2 (PV_G2), which is encoded by the *Cer* retrovirus *envelope* gene (Fig. 2B). We used PV_G2 rather than PV_G1 in our searches because the former was much more conserved.

Interestingly, no matches could be found with even remote similarity (e.g., $E <$ 0.1) to PV_G2 in any of the genome sequences evaluated from clades 2 and 8 (Fig. 1; Table 1). However, highly significant matches (e.g., $E <$ 1E-20) were found in the genome sequences of all clades 9, 10 and 12 nematodes evaluated (Fig. 1; Table 1). Noteworthy, we observed extensive variation in the total number of both fragmented and complete *Cer* retroviruses (Table 1), but we cannot rule out the possibility that this finding was simply a consequence of incomplete genome sequences. Interestingly, complete *Cer* retroviruses with identifiable 5'- and 3'-LTRs, which might indicate transcriptional activity (Xu and Wang 2007), were found in almost all clades 9 and 12 nematodes (9/11; Table 1). Taken together, these results indicated that *Cer* retroviruses likely appeared in the genome of a common ancestor of clades 9 through 12 nematodes, that they remain present in multiple nematode species from these clades, and that they may still be infectious, including in clade 12 PPN.

**Table 1. Summary of Cer retroviruses in nematode genomes**

| Nematode | Clade | Lifestyle | No. of *envelope* genes | No. of *Cer* retroviruses | No. of *Cer* retroviruses w/ LTRs |
|---|---|---|---|---|---|
| *G. pallida* | 12 | Plant parasitic | 13 | 8 | 4 |
| *H. glycines* | 12 | Plant parasitic | 65 | 31 | 14 |
| *M. incognita* | 12 | Plant parasitic | 29 | 19 | 14 |
| *M. hapla* | 12 | Plant parasitic | 10 | 3 | 2 |
| *B. xylophilus* | 10 | [a]Plant parasitic | 6 | 2 | 0 |
| *C. briggsae* | 9 | Free-living | 4 | 2 | 0 |
| *C. elegans* | 9 | Free-living | 3 | 2 | 2 |
| *C. remanei* | 9 | Free-living | 21 | 11 | 6 |
| *P. pacificus* | 9 | Free-living | 6 | 0 | 0 |
| *N. americanus* | 9 | Human parasitic | 26 | 12 | 5 |
| *H. contortus* | 9 | Animal parasitic | 42 | 15 | 4 |
| *H. bacteriophora* | 9 | Insect parasitic | 28 | 10 | 2 |
| *A. suum* | 8 | Animal parasitic | 0 | n/a | n/a |
| *B. malayi* | 8 | Human parasitic | 0 | n/a | n/a |
| *L. loa* | 8 | Human parasitic | 0 | n/a | n/a |
| *T. spiralis* | 2 | Animal parasitic | 0 | n/a | n/a |
| *T. suis* | 2 | Animal parasitic | 0 | n/a | n/a |
| *T. trichiura* | 2 | Human parasitic | 0 | n/a | n/a |

[a] *Bursaphelenchus xylophilus* is a free-living fungivore and facultative plant parasite.

## 4.4 *Cer* retroviruses are significantly associated with known HGT genes in genomes of clade 12 PPN

Next, we determined whether previously documented HGT genes in the genomes of clade 12 PPN are associated with *Cer* retroviruses, which could potentially reflect their contribution to HGT. We used tblastn searches for known HGT genes in the genome sequences of the four clade 12 PPN *Heterodera glycines*, *Globodera pallida*, *Meloidogyne incognita* and *M. hapla*. We had to perform this initial search because the conservation of some of the HGT genes in these PPN was unknown. Results from these searches are summarized in Table 2. Once the HGT genes were identified within the genomic scaffolds, we performed blastx searches of the DNA sequences that flanked each HGT gene within each scaffold. Depending on the size of each genomic scaffold,

around 10 to 20 genes were examined in both the 5' and 3' directions from each HGT gene locus. Most of our searches focused on the genome sequences from cyst nematodes *H. glycines* and *G. pallida* because we found that these scaffolds were much larger compared to the root-knot nematode scaffolds (see Materials and Methods for details).

Remarkably, eight of the seventeen HGT gene families analyzed resulted in at least one member that was found in close proximity to either a complete or fragmented *Cer* retrovirus (Fig. 3; Table 2). For example, two complete *Cer* retroviruses were found between the *Hg-GLAND13 invertase* and two *expansins* with five and zero intervening genes, respectively (Fig. 3). A *chorismate mutase* in *G. pallida*, a *polyglutamate synthase* in *H. glycines*, and *thiM* in *G. pallida* were associated with either complete or fragmented *Cer* retroviruses with zero or a maximum of three intervening genes (Fig. 3). Although relatively distant, but still within the maximum 20 gene limit in our searches, another *invertase* and *expansin*, and both *tenA* and *NodL* in *H. glycines* were found in proximity to either complete or fragmented *Cer* retroviruses (Fig. 3). Moreover, in a separate tblastn search of the draft *M. incognita* genome sequence CNS 2007-10 at the INRA web portal we found co-localization of a *pectate lyase* with a *Cer* retrovirus in scaffold MiV1ctg171 (Table 2). However, we could not obtain nucleotide sequences from this genome resource for further inspection.

Furthermore, we performed blastx searches on more than 100 randomly selected scaffolds from both *H. glycines* and *G. pallida* genome sequences. These searches resulted in zero complete or fragmented *Cer* retroviruses in close proximity to randomly selected control genes within our selected 20-gene limit (examples are included in Table S1). The latter findings, in addition to the relatively low number of total *Cer* retroviruses

**Table 2. Clade 12 PPN HGT genes evaluated for co-localization with *Cer* retroviruses**

| [a]Name | [b]Accession | Presence in PPN genomes | | | | Co-localized with *Cer* retroviruses | |
|---|---|---|---|---|---|---|---|
| | | [f]*Hg* | *Gp* | *Mi* | *Mh* | *Hg* | *Gp* |
| *Plant cell-wall modification* | | | | | | | |
| GH5 cellulase | AAC48327.1 | ✔ | ✔ | ✔ | ✔ | | |
| expansin | ADY02960.1 | ✔ | ✔ | ✔ | ✔ | ✔ | |
| [c]pectase lyase | AAF80747.1 | ✔ | ✔ | ✔ | ✔ | | |
| *Modulation of plant defenses* | | | | | | | |
| chorismate mutase | AEA07500.1 | ✔ | ✔ | ✔ | ✔ | | ✔ |
| *Nutrient processing* | | | | | | | |
| [d]thiD (B1) | ACZ34279.1 | ✔ | ✔ | | | | |
| thiE (B1) | ACZ34278.1 | ✔ | | | | | |
| thi4 (B1) | ACZ34282.1 | ✔ | ✔ | | | | |
| thiM (B1) | ACZ34284.1 | ✔ | ✔ | | | | ✔ |
| tenA (B1) | ACZ34283.1 | ✔ | ✔ | | | ✔ | |
| panC (B5) | ACZ34280.1 | ✔ | ✔ | ✔ | ✔ | | |
| SNO (B6) | ACF10393.1 | ✔ | ✔ | | | | |
| SNZ (B6) | ACF10392.1 | ✔ | ✔ | | | | |
| bioB (B7) | ACZ34281.1 | ✔ | ✔ | ✔ | ✔ | | |
| GH32 invertase | KJ825724.1 | ✔ | ✔ | ✔ | ✔ | ✔ | |
| Polyglut synthase | CAC84452.1 | ✔ | ✔ | | | ✔ | |
| *Feeding strategy* | | | | | | | |
| NodL | [e]CAA35590.1 | ✔ | ✔ | ✔ | ✔ | ✔ | |
| *Unknown process* | | | | | | | |
| GLAND1 | KJ825712.1 | ✔ | ✔ | | | | |

[a] HGT genes from the clade 12 PPN were selected according to whether they were found in the *H. glycines* genome sequence.

[b] Accession indicates a representative of the respective HGT gene family that was used to search each clade 12 PPN genome sequence.

[c] We identified co-localization of *pectate lyase* with PV_G2 in the *Meloidogyne incognita* draft genome assembly CNS 2007-10, but were unable to inspect the scaffold sequence.

[d] Protein names thiD through bioB contain in parentheses the vitamin B pathways for which they are involved in, according to (Craig et al. 2008; Craig et al. 2009).

[e] We were unable to obtain published NodL sequences from PPN, and thus used a protein sequence from *Rhizobium leguminosarum* (among the suspected donor bacteria) as query.

[f] *Hg*, *H. glycines*; *Gp*, *Globodera pallida*; *Mi*, *M. incognita*; *Mh*, *Meloidogyne hapla*.

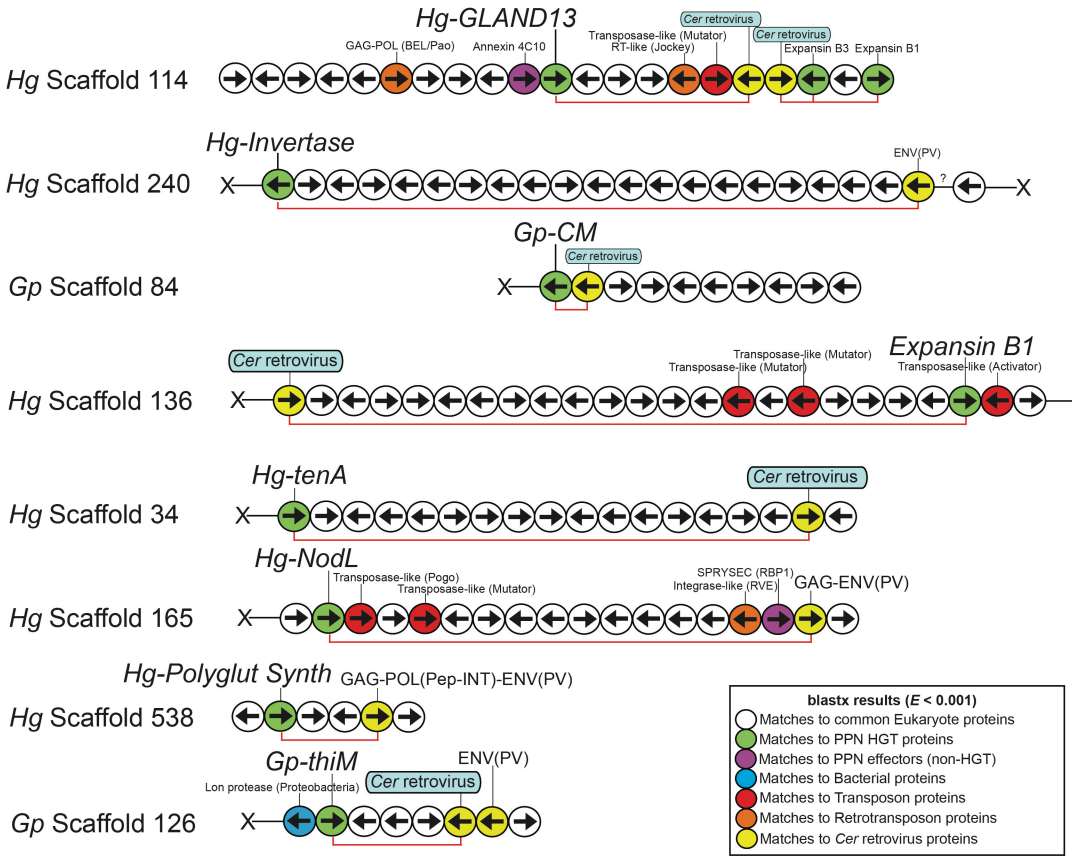



**Fig. 3. *Cer* retroviruses are in close proximity to HGT genes in the genomes of cyst nematodes *H. glycines* and *G. pallida*.** Scaffold numbers from *H. glycines* (*Hg*) and *G. pallida* (*Gp*) genome sequences are indicated to the left. Each circle corresponds to a single protein-coding gene. Color designations for each protein-coding gene are described in the box in the lower right-hand corner. Arrows within the circles indicate the relative orientations of each protein-coding gene. Red connecting lines illustrate close proximity of HGT genes (i.e., within our 20 gene limit) with fragmented or complete *Cer* retroviruses. Complete *Cer* retroviruses are highlighted teal. For fragmented *Cer* retroviruses, the partial elements detected are listed as follows: ENV(PV), Phlebovirus *envelope* gene; GAG-ENV(PV), *Cer* retrovirus missing the entire *pol* gene; GAG-POL(PEP-INT)-ENV(PV), *Cer* retrovirus missing the reverse transcriptase within the *pol* gene (see Fig. 2). Terminal ends of the genomic scaffolds are illustrated with an X. Regions within scaffolds with low-quality sequence are indicated with a question mark above. RT, reverse transcriptase. Descriptions of all protein-coding genes are provided in Table S1.

that are distributed throughout the entire genome sequences (Table 1), indicated a P-value much less than 0.01 for the association of HGT genes with *Cer* retroviruses. Taken

together, these results indicated a significant genomic association of *Cer* retroviruses with HGT genes in the analyzed genomes of clade 12 PPN.

During our searches for *Cer* retroviruses, we also identified in close proximity to some HGT genes protein coding sequences that matched to the following TE families: retrotransposons *Jockey*, *BEL/Pao* and *Ty3/Gypsy* (i.e., *Cer* retrotransposons), and DNA transposons *Activator*, *Mariner*, *Mutator*, *Pogo* and *Helitron* (e.g., see Fig. 3; Table S1). However, we identified similar patterns of TEs also in close proximity to random control genes in the majority of the more than 100 randomly evaluated scaffolds, negating any significance for the observed close proximity of TEs with the HGT genes (examples are included in Table S1). The latter findings are consistent with (Danchin et al. 2010). Therefore, our results do not indicate an association of HGT genes with TEs, per se, but rather, specifically to *Cer* retroviruses in the genome sequences of clade 12 PPN.

Duplication and genome rearrangement of the originally acquired HGT genes can explain why not all PPN HGT genes are found in close proximity to *Cer* retroviruses. Following an HGT event, the acquired gene experiences positive selection for the amelioration of donor genome features, in some cases insertion of spliceosomal introns, and might evolve a completely different function in the recipient genome (Scholl et al. 2003; Huang 2013; Wijayawardena et al. 2013; Noon and Baum, manuscript submitted). All of these changes might involve genome rearrangements. Thus, it is quite remarkable that eight of the seventeen HGT gene families evaluated (47%) resulted in at least one member (possibly the originally acquired gene) in close proximity to either complete or fragmented *Cer* retroviruses. Once the list of sequenced genomes for clade 12 PPN

grows, additional, known HGT genes, and possibly newly identified HGT genes, may or may not be found in close proximity to *Cer* retroviruses.

In conclusion, we present the following hypothesized mechanism for HGT in clade 12 PPN. *Cer* retroviruses, which acquired infectious abilities via acquisition of a Phlebovirus *envelope* gene (Malik et al. 2000; Ganko et al. 2001) most likely in a common ancestor of clades 9 through 12 nematodes, may have allowed for the 'capture' of foreign DNA in cells throughout the nematode body. Homology-independent, intermolecular strand-transfer is a plausible molecular mechanism for which the foreign DNA could be 'captured' by *Cer* retroviruses (Varmus and Brown 1989; Malik et al. 2000). Then, the *Cer* retrovirus carrying the foreign genetic material spreads intercellularly, enters the germ line, and then the *Cer* retrovirus' integrase provides the molecular mechanism for integration of the foreign DNA into the genomes of germ cells thus providing the foundation for vertical transmission. The acquired gene(s) could then experience rearrangements and positive selection to be functional once again in the new genome. This hypothesized mechanism would thereby enable vertical transmission given a selective advantage of the acquired gene(s). Furthermore, since *Cer* retroviruses most likely appeared only in an ancestor of clades 9 through 12 nematodes, and is thus likely absent from all other organisms, we also hypothesize that divergent eukaryotes may have different, highly evolved mechanisms for HGT. These potentially different HGT mechanisms may or may not involve retroviruses. Future studies will hopefully show whether *Cer* retroviruses are present in the germ cells of clades 9 through 12 nematodes, in particular clade 12 PPN, similar to the closely related *Cer1* retrotransposon in *C. elegans*. Also, these studies may lead to developments for testing whether *Cer*

retroviruses were (and maybe still are) the mechanistic agents for HGT in these nematodes.

## 4.5 Materials and Methods

### 4.5.1 Searches of nematode genome sequences

Tblastn searches (Altschul et al. 1997) were performed against the nematode genome sequences listed in Table S2, Apendix C. As query, we implemented glycoprotein G2 from Uukuniemi virus (GenBank accession NP_941986.1) as the representative for PV_G2, and used a relatively sensitive E-value threshold equal to 0.1. For our searches of HGT genes in clade 12 PPN, we used a relatively stringent E-value threshold of 1E-10.

### 4.5.2 Scans for protein-coding genes in the genome sequences of clade 12 PPN

To identify protein-coding genes near HGT genes in the genome sequences of clade 12 PPN, we used homology-based searches. Blastx (Altschul et al. 1997) scans were run with E-value thresholds equal to 0.001 on the genomic DNA sequences that flanked the HGT genes in the respective scaffolds of the *H. glycines* and *G. pallida* draft genome assemblies. We searched *H. glycines* and *G. pallida* genome assemblies specifically because the scaffolds were found to be much larger than for the other clade 12 PPN analyzed, especially the scaffolds containing PV_G2 (e.g., *M. incognita* μ-'scaffold'-length = 24.2-Kbp, *G. pallida* μ-scaffold-length = 188.7-Kbp and *H. glycines* μ-scaffold-length = 193.4-Kbp). This allowed us to identify substantially more protein-

coding genes in the scaffolds of these two cyst nematodes, and to identify co-localization of their HGT genes with *Cer* retroviruses (Fig. 3; Table 1). We arbitrarily set a 20 gene limit for our blastx scans on the DNA sequences that flanked each HGT gene, both 5' and 3'. However, the HGT genes were commonly positioned near the ends of the scaffolds, so in those cases we were limited (e.g., see Fig. 3).

### 4.5.3 LTR predictions for complete *Cer* retroviruses

Nematode genomic DNA sequences that significantly matched to PV_G2 were conceptually translated and inspected for matches to additional *Cer* retroviral *gag*, *pol* and *envelope* protein domains (see Fig. 2B) using blastp, CD-search (Marchler-Bauer and Bryant 2004; Marchler-Bauer et al. 2015) and InterProScan 5 (Jones et al. 2014). 10-Kbp upstream and downstream from complete *Cer* retroviruses were subjected to full-length LTR retrotransposon predictions using the software LTR_FINDER (Xu and Wang 2007). In every analysis, default parameters were used except that we implemented the *C. elegans* tRNA database for PBS predictions.

## 4.6 Authors' contributions

JBN designed and performed all analyses, and wrote the manuscript. TJB supervised the work and co-wrote the manuscript with JBN.

# CHAPTER 5. A *PLASMODIUM*-LIKE VIRULENCE EFFECTOR OF THE SOYBEAN CYST NEMATODE SUPPRESSES PLANT INNATE IMMUNITY

A paper submitted to *New Phytologist*

Jason B. Noon, Mingsheng Qi, Danielle N. Sill, Usha Muppirala, Sebastian Eves-van den Akker, Thomas Maier, Drena Dobbs, Melissa Mitchum, Tarek Hewezi, Thomas J. Baum

## Summary


- *Heterodera glycines*, the soybean cyst nematode, delivers effector proteins into soybean roots to initiate and maintain an obligate parasitic relationship. *HgGLAND18* encodes a candidate *H. glycines* effector and is strongly expressed throughout the infection process.
- We used a combination of molecular, genetic, bioinformatic and phylogenetic analyses to determine the role of *HgGLAND18* during *H. glycines* infection.
- *HgGLAND18* is necessary for virulence in compatible interactions with soybean. The encoded effector strongly suppresses both PTI and ETI innate immune responses, and immunosuppression requires the presence and coordination between multiple protein domains. The N-terminal domain in HgGLAND18 contains unique sequence similarity to domains of an immunosuppressive effector of *Plasmodium* spp., the malaria parasites. The *Plasmodium* effector domains functionally complement the loss of the N-terminal domain from HgGLAND18.

- In-depth sequence searches and phylogenetic analyses demonstrate convergent evolution between effectors from divergent parasites of plants and animals as the cause of sequence and functional similarity.

## 5.1 Introduction

*Heterodera glycines*, the soybean cyst nematode, is an economically important, obligate biotroph of soybean that feeds only during its sedentary life stage. These sedentary nematodes are completely reliant on the reprogramming and survival of specialized feeding cells whose formation they induce in soybean roots.

*H. glycines* produces effector proteins that contain N-terminal secretion signal peptides and are released into the plant via a mouthspear (Mitchum et al. 2013). More than eighty distinct *H. glycines* effectors have been documented so far (Gao et al. 2001; Wang et al. 2001; Gao et al. 2003; Noon et al. 2015). *Heterodera* cyst nematode effector characterizations implicate these proteins in cell wall modifications (Hewezi et al. 2008a), auxin transport and signaling (Lee et al. 2011; Hewezi et al. 2015), polyamine metabolism (Hewezi et al. 2010), ubiquitination (Tytgat et al. 2004) and mimicry of regulatory peptides (Wang et al. 2010; 2011). Furthermore, cyst nematode effectors have been implicated in the suppression or activation of plant innate immunity [reviewed in (Hewezi and Baum 2013; Mitchum et al. 2013; Goverse and Smant 2014; Hewezi 2015)].

The plant innate immune system consists of basal surveillance systems and a wide spectrum of defense mechanisms including a hypersensitive cell death response (HR). Microbes contain microbe-associated molecular patterns (MAMPs) that are recognized

by plant extracellular pattern-recognition receptors (PRRs). MAMP recognition by a PRR induces pattern-triggered immunity (PTI), the basal immune response. As an evolutionary consequence, many pathogen effectors suppress PTI, which in turn drove the evolution of plant resistance (*R*) genes that detect the presence of effectors and trigger massive defense mechanisms, so-called effector-triggered immunity (ETI). In general, PTI and ETI involve similar salicylic acid (SA)-responsive immune signaling, with the latter having a much stronger output that results in HR (Jones and Dangl 2006; Spoel and Dong 2012; Newman et al. 2013). Plant-parasitic nematodes contain MAMPs that induce PTI (Manosalva et al. 2015) and effectors that trigger ETI (Goverse and Smant 2014).

The recently identified *H. glycines* putative effector *HgGLAND18* is expressed specifically in the dorsal gland cell during parasitism, and the encoded effector sequence has no detectable homologs in the non-redundant database (nr) at E-value < 0.001 (Noon et al. 2015). Here, we describe the functional characterization of HgGLAND18 using a combination of molecular, genetic, bioinformatic and phylogenetic analyses. We determine that expression of *HgGLAND18* is necessary for full *H. glycines* virulence and that the encoded effector suppresses both PTI and ETI. Additionally, we determine that the HgGLAND18 immunosuppressive function is not conditioned by a single discrete protein domain but requires the presence and coordination of different protein regions. Bioinformatic and phylogenetic analyses revealed significant sequence similarity between an N-terminal region of HgGLAND18 and specific protein domains (RI, RR and RII+) of the immunosuppressive circumsporozoite protein (CSP) effector of *Plasmodium* spp., the malaria parasites.

Animal innate immune systems are likewise targeted by pathogen effectors (Espinosa and Alfano 2004) and *Plasmodium* CSP is one such example. All CSPs contain seven distinct protein domains [signal peptide, PEXEL/VTS motifs, region I (RI), a species-specific and immunodominant tandem repeat region (RR), region III (RIII), region II+ (RII+) and a glycosylphosphatidylinositol (GPI)-anchor for attachment of CSP to the sporozoite surface] that delineate different functions (Fig. S1, Appendix D) (Coppi et al. 2011). CSP assists in both the migration to and entry into liver cells (Coppi et al. 2011), and this entry involves coordinated binding of RIII and RII+ domains to an extracellular surface ligand (Coppi et al. 2011). After sporozoite entry into liver cells the parasite is encapsulated by the parasitophorous vacuole membrane (PVM) (Graewe et al. 2012). PEXEL/VTS motifs are required for effector translocation through the PVM (Singh et al. 2007). In rodent malarias, CSP enters liver cells and binds to importin-α proteins via a nuclear localization signal (NLS) within the RII+ domain (Singh et al. 2007). This interaction outcompetes NFκB for nuclear uptake, thereby inhibiting the innate immune response (Singh et al. 2007). Furthermore, in older reports, *Plasmodium falciparum* CSP was shown to enter and kill immune cells by inhibiting protein synthesis most likely from the RNA-binding properties of domains RI, RR and RII+ (Hugel et al. 1996; Frevert et al. 1998). Thus, *Plasmodium* CSPs are potent immunosuppressors in animal cells when delivered into the cytoplasm, and the effector function heavily relies on domains RI, RR and RII+.

Extensive database searches determined that the similarity between HgGLAND18 and the *Plasmodium* CSPs is unlikely to be found in proteins from other organisms, and thus, in combination with additional data, cannot be explained by homology and

divergent evolution. Furthermore, we show that deletion of the N-terminal region from HgGLAND18 abolishes immunosuppression, but remarkably, *Plasmodium* CSP domains are able to fully complement the function of the HgGLAND18 deletion mutants. We conclude that the observed sequence similarities between HgGLAND18 and the requisite *Plasmodium* CSP domains is best explained by convergence due to similar immunosuppressive functions in their respective host cells.

## 5.2 Materials and Methods

### 5.2.1 Nematodes and plants

*H. glycines* were propagated on soybean according to (Niblack et al. 1994), *Heterodera schachtii* on sugar beet, and *Meloidogyne incognita* on tomato at Iowa State University. The soybean lines used in this study were obtained from the USDA Soybean Germplasm Collection, University of Illinois. *Nicotiana benthamiana* were grown in a growth chamber at 25˚C with 16:8-hr light/dark cycles.

### 5.2.2 RNA isolation and cDNA synthesis

Nematodes were isolated from roots by macerating in a blender followed by sieving and separation on a sucrose gradient, were frozen, and homogenized with sterile 1.0-mm diameter Zirconia Beads (BioSpec) in a Mini-BeadBeater (BioSpec). Frozen plant tissues were homogenized with sterile 3.5-mm diameter Glass Beads (BioSpec). Total RNA was isolated with the NucleoSpin Kit (Clontech). Yields and integrity were

assessed using a NanoDrop and agarose gel electrophoresis, respectively. cDNA synthesis was performed with qScript (Quanta).

### 5.2.3 RT-PCR and isolation of *HgGLAND18* cDNA sequences

RT-PCR was performed with *Taq* Polymerase (NEB). For RT-PCR on soybean cDNA, *GmPolyubiquitin3* (GenBank: D28123.1) was used as reference. For RT-PCR on *H. glycines* cDNA, *HgActin1* (GenBank: AF318603.2) was used as reference. TrackIt 10-bp DNA Ladder (Invitrogen) was used for RT-PCR analysis of *HgGLAND18* isoforms. *HgGLAND18* cDNAs were isolated with Platinum *Taq* (Invitrogen) for PCR, and purified products were ligated into pGEM-T Easy (Promega) and sequenced at Iowa State University.

### 5.2.4 Isolation of *HgGLAND18* genomic sequences

Genomic DNA was isolated from both homogenized nematode egg and soybean leaf tissues according to (Blin and Stafford 1976). Yields and integrity were assessed as described above. PCR was performed on *H. glycines* genomic DNA with Platinum *Taq*, and purified DNA was ligated into pCR-XL-TOPO using the TOPO XL Kit (Invitrogen). Sequencing by primer walking was performed at Iowa State University.

### 5.2.5 Hairy root RNAi experiments

Nucleotides 84-546 were PCR-amplified with Platinum *Taq* from an HgGLAND18 (isoform 3-2) coding DNA sequence (CDS) pGEM-T Easy clone. PCR products were restriction-digested with AscI and SwaI (NEB) for the sense fragment, and

AvrII and BamHI (NEB) for the antisense fragment, cloned into pG2RNAi2 (GenBank: KT954097) and sequenced as above. Transgenic hairy roots were generated and nematode infection assays were performed similar to (Liu et al. 2012), only experiments were setup in 6-well plates with randomization, as in (Baum et al. 2000). Statistical differences were tested using the t-test in JMP Pro 11.

### 5.2.6 Ectopic expression in hairy roots

Nucleotides 40-546 were PCR-amplified with Platinum *Taq* from an HgGLAND18-3-2 CDS pGEM-T Easy clone. The PCR product was restriction-digested with SwaI and BamHI, cloned into pG2XPRESS and sequenced as above. pG2XPRESS was derived from pG2RNAi2; the *GUS* linker sequence was digested out. Transgenic hairy roots were generated as above.

### 5.2.7 Assessment of growth defects

Growth rate was measured as the inverse of the number of days that parent roots took to fill an entire plate after transfer ($n = 5$). Biomass was measured as the percentage of dry root weight with the vector control mean set to 100% ($n = 5$).

### 5.2.8 qRT-PCR

One-step qRT-PCR was performed with qScript One-Step qRT-PCR Kit (Quanta). 10-ng of total RNA was used as template. Protocol: 49˚C for 10-min, 95˚C for 5-min, 35 cycles of 95˚C for 15-sec and 60˚C for 45-sec. Minus RT reactions were always included. *HgActin1* was used as calibrator. Data were analyzed using the $2^{-\Delta\Delta CT}$

method (Livak and Schmittgen 2001), and statistical differences were tested using the t-test in JMP Pro 11. Two-step qRT-PCR was performed using iQ SYBR Green Supermix (Bio-Rad). 1-µg of total RNA was used for cDNA syntheses, cDNA samples were diluted to 40-µL, and 1-µL of cDNA was used as template. Protocol: 95˚C for 3-min, 40 cycles of 95˚C for 15-sec and 60˚C for 30-sec. The same estimated amount of total RNA was always included for each cDNA sample. *NbActin1* (GenBank: AY594294.1) was used as calibrator. Data were analyzed as above, and statistical differences were tested using the Tukey-Kramer HSD test in JMP Pro 11. In every qRT-PCR experiment, 3 biological and 4 technical replicates were used. Amplification specificities were verified by melting curve analysis and agarose gel electrophoresis. Melting curve analysis protocol: 95˚C for 1-min, 55˚C for 10-sec and a slow temperature ramp from 55-95˚C. qRT-PCR was performed on an iCycler iQ Real-Time PCR Detection System (Bio-Rad).

### 5.2.9 Insertion and deletion mutagenesis

Insertion and deletion mutagenesis was performed with overlap-extension PCR (Ho et al. 1989). For *HgGLAND18* mutants, an HgGLAND18-3-2 CDS pGEM-T Easy clone was used as template. To generate the chimeric fusion proteins for *Plasmodium fieldi* CSP, a synthetic clone was ordered from GenScript and used as template.

### 5.2.10 Southern blot

gDNA samples were treated with RNase H (Invitrogen). 10-µg of gDNA was restriction-digested overnight with EcoRI and HindIII (Invitrogen) separately. DNA

transfer, probe hybridization and signal detection were performed according to (Hewezi et al. 2006).

### 5.2.11 PTI and ETI suppression experiments

PCR products for wild-type *HgGLAND18*$^{-sp}$ and mutants were TOPO-cloned into pENTR with the pENTR/D-TOPO Kit (Invitrogen). pENTR clones were gateway-cloned into pEDV6 (Fabro et al. 2011) with LR Clonase (Invitrogen), and sequenced as above. Tri-parental mating was used for conjugation of pEDV6 vectors into *Pseudomonas fluorescens* strain EtHAn and *Pseudomonas syringae* pathovar *tomato* strain DC3000. PTI and ETI suppression experiments were performed as in (Chakravarthy et al. 2009). Bacteria were suspended in 10-mM $MgCl_2$ and infiltrated into *N. benthamiana* leaves with OD600s equal to 0.2 and 0.02, respectively. For quantitative qRT-PCR experiments bacteria were infiltrated into entire *N. benthamiana* leaves.

### 5.2.12 Protein secretion assays

Accumulation of AvrRPS4:HA:HgGLAND18$^{-sp}$ in *Pseudomonas* and its secretion by the type III secretion system was verified according to (Fabro et al. 2011). Pellet and supernatant fractions were analyzed by SDS–PAGE, electro blotted onto PVDF membrane (Bio-Rad), and probed with anti–HA–HRP antibody (Roche). Bands were visualized using PICO kit (Thermo) and imaged with Kodak scientific imaging film.

### 5.2.13 NCBI database searches

RR sequences from eighteen *Plasmodium* CSP sequences (Table S1, Appendix D), as well as the HgGLAND18 (isoform 3-5) repeats were searched against every NCBI database with DELTA-BLAST (Boratyn et al. 2012) using a sensitive E-value threshold of 10. All hits were collected into FASTA files. An automated bioinformatics pipeline was generated that screened for tandem repeats with Internal Repeats Finder (IRF) (Pellegrini et al. 1999), repeat size with TRUST (Szklarczyk and Heringa 2004) and our own script was written to extract the tandem repeats from each hit. Any hits that did not match the tandem repeat structure of each *Plasmodium* CSP RR or the HgGLAND18 repeats were removed. Bl2seq was then used to eliminate hits that did not contain tandem repeats with similar sequences (i.e., E-value > 1.0). All hits were then evaluated for precisely paired repeats (see Fig. 6A,B). Survivors were then blastp-searched against each *Plasmodium* sp. nr database using both standard and sensitive parameters (i.e., word size = 2, BLOSUM45, no adjustments) with E-value thresholds of 1000, and were inspected manually with Multalin (Corpet 1988), to search for additional alignment to RI and RII+. In separate searches, multiple sequence alignments (MSAs) of all eighteen *Plasmodium* RI and RII+ sequences were generated with MUSCLE (Edgar 2004) and were submitted to HMMER3 (Eddy 1998) using standard parameters. Each profile-hidden Markov model was searched against all NCBI databases, and hits were collected into FASTA files. All hits were screened for additional domains as performed above. Finally, every protein in NCBI databases that was found to contain a CSP-like identifier, which we considered possible homologs, was also run through our screens, none of which survived.

### 5.2.14 Nematode database searches

Tblastn-searches were performed against all nematode genomic and transcriptomic sequences at Nematode.net (Wylie et al. 2004), and raw sequence reads from eight plant-parasitic nematode species (Table S3, Appendix D) with HgGLAND18 as query. In general, our searches used E-value thresholds of 0.001, and additional searches were performed with more sensitive thresholds but the resulting hits aligned only randomly with HgGLAND18, and thus, these hits were discarded. Noteworthy, the combination of nematode sequences from Nematode.net and the raw sequence reads covered the major lineages of the plant-parasitic nematode suborder Hoplolaimina (Holterman et al. 2006).

### 5.2.15 Model selection analyses

Model selection analysis assesses the likelihoods of different models of sequence evolution (Theobald 2010). In our analyses, Bayesian and corrected Akaike Information Criteria were used as scores (Tamura et al. 2011). By statistical convention, a score difference of greater than 5 is strong empirical evidence for the better model (Burnham et al. 1998; Theobald 2010). Four control sequences were included in the analysis. The first two controls were HgGLAND8 and the *B. cereus* 'circumsporozoite protein', which were the top nr blastp hits for HgGLAND18. The third control was human SARMP2 (GenBank: XP_006714000), which was the top nr blastp hit for the three *Plasmodium* CSPs in question. The fourth control was *Plasmodium falciparum* EMP1 (GenBank: AEA03008), which was a sequence in *Plasmodium* not related to *Plasmodium* CSPs. MSAs were generated via MUSCLE within MEGA6 (Tamura et al. 2013), and poorly

aligned regions were removed. Model selection analysis was performed in MEGA6 on each MSA. For model selection, different tree topologies (i.e., evolutionary models) were generated with the Topology Editor tool within MEGA6. Each model selection analysis was repeated at least once with identical results.

### 5.2.16 Phylogenetic analyses

Phylogenetic trees were constructed in MEGA6 with bootstrapped Maximum Likelihood estimation with the best-scoring model of amino acids substitution that resulted from model selection analyses. 100 bootstrap replications were used. Reported are the best-scoring ML phylogenetic trees with bootstrap values indicated on the corresponding nodes.

## 5.3 Results

### 5.3.1 *HgGLAND18* contains a polymorphic tandem repeat region

We previously reported the HgGLAND18 sequence (GenBank: KJ825729.1) obtained from a draft genome that was sequenced from an inbred *H. glycines* population (line TN10; Noon et al. 2015). The TN10 allele of *HgGLAND18* contains eight exons, and exon 2 is very small encoding only 11 amino acids (Noon et al. 2015). To explore HgGLAND18 coding sequence variability, we performed RT-PCR using RNA obtained from a mixture of life stages from an outbred *H. glycines* field population. High-resolving agarose gel electrophoresis revealed six distinct bands of 110- to 270-bp (Fig. 1A). Subsequent sequencing of 30 different clones derived from these amplification products

revealed that the observed size differences were due to two main sequence polymorphisms. One, HgGLAND18 amplification products fell into four different sequence groups depending on the absence/presence of a single amino acid (aa) codon (N) close to the N-terminus or a group of three aa codons (VNG) towards the center of the protein. These sequence groups likely correspond to allelic variation or may even indicate the presence of a gene family since multiple intense bands were found in a Southern blot of genomic DNA obtained from another inbred *H. glycines* population (line OP50; Fig. S2, Appendix D). We named these four sequence types HgGLAND18 isoform 1 through 4 (Fig. 1B). Second, we discovered that HgGLAND18 contains a tandem repeat region in the N-terminal half and that within the four HgGLAND18 isoforms mentioned above, clones differed in the number (0-5) of tandem repeats (Fig. 1B). We added a number designator to each isoform name to indicate the number of repeats present. Noteworthy, isoform 3 with 2 repeats (HgGLAND18-3-2; GenBank KT954103) was substantially overrepresented (22/30 clones) in the sequencing (Fig. 1C). Interestingly, we found that each repeat actually corresponds to exon 2 from the TN10 allele (Fig. 1D). Moreover, we obtained genomic DNA clones of *HgGLAND18* from inbred line OP50 and found that compared to the TN10 allele, exon 2 is duplicated to form a tandem repeat (Fig. 1D). These findings indicate that there are variable numbers of *HgGLAND18* repeats between, and within, at least some *H. glycines* populations.

We also assessed the developmental expression patterns of *HgGLAND18* in the six *H. glycines* life stages separately (i.e., egg to adult female) of the field population by RT-PCR followed by sequencing of amplification products. Consistent with cloning

efficiency, HgGLAND18-3-2 was by far the most abundant transcript in all *H. glycines* life stages and showed similar intensity throughout the life cycle (Fig. 1E).

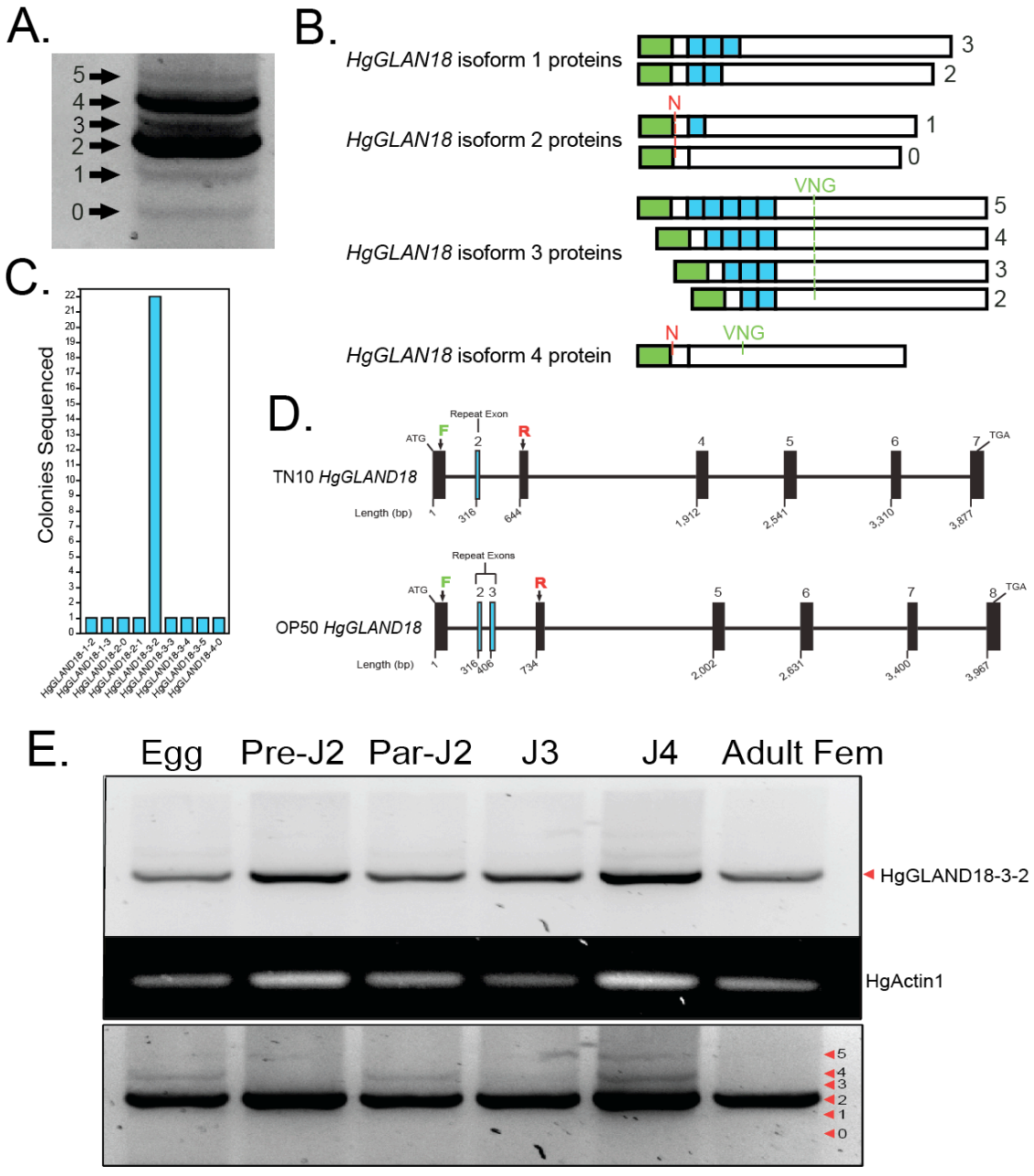



**Fig. 1. A single *HgGLAND18* isoform predominates throughout the *Hetereodera glycines* life cycle. (A)** RT-PCR on the *HgGLAND18* tandem repeat region using mixed parasitic *H. glycines* life stages. Bands are labeled according to the number of repeats. Shown is an inverted gel image. **(B,C)** RT-PCR was performed on the full-length

HgGLAND18 coding DNA sequence using mixed parasitic *H. glycines* life stages, and a single, smeared band was cloned, and plasmids obtained from 30 different bacterial colonies were sequenced. (B) Four different *HgGLAND18* isoforms were identified from codon insertions/deletions labeled at the corresponding positions. Multiple proteins from each isoform are shown with the repeats colored light blue (signal peptide is colored green). (C) Number of colonies that resulted in each *HgGLAND18* isoform (22/30 colonies = HgGLAND18-3-2). **(D)** *HgGLAND18* gene structures in *H. glycines* lines TN10 and OP50. TN10 *HgGLAND18* was obtained from a *H. glycines* genome sequence (Noon et al. 2015) and OP50 *HgGLAND18* was PCR-amplified from genomic DNA, cloned and sequenced. Exons and introns are illustrated as boxes and horizontal lines, respectively. A scale of nucleotide positions is provided below each *HgGLAND18* gene. Exons that encode individual repeats are colored light blue and labeled. Annealing sites for the primers used in panel (A) are shown within the corresponding exons. **(E)** RT-PCR on the *HgGLAND18* tandem repeat region as in panel (A) on each individual stage of the *H. glycines* life cycle, with *HgActin1* as reference. Top, inverted gel image of *HgGLAND18*; middle, regular gel image of *HgActin1*; bottom, inverted gel image of *HgGLAND18* with greater exposure. Bottom, bands are labeled according to the number of repeats. Top, the most intense band was purified from each lane and sequenced, which resulted exclusively in the HgGLAND18-3-2 isoform.

### 5.3.2 Host-induced RNAi of *HgGLAND18* decreases *H. glycines* virulence

To determine the importance of *HgGLAND18* for *H. glycines* infection, we performed host-induced RNA interference (RNAi) to post-transcriptionally silence *HgGLAND18* in the nematodes in hairy root assays. A hairpin construct was generated to target nucleotides (nt) 84-546 of the *HgGLAND18* gene (*HgGLAND18i*; Fig. 2A), which was placed under transcriptional control of a soybean *polyubiqutin* promoter [GenBank: EU310508.1; (Hernandez-Garcia et al. 2009)]. Noteworthy, the targeted region of *HgGLAND18* was pre-determined through blastn-searches to be absent from soybean and to match only *HgGLAND18* in the *H. glycines* genome at E-value < 1.0.

Our T-DNA construct also contained a functional *GFP* gene, which allowed the identification of transgenic soybean roots by GFP expression. RT-PCR analyses determined transgenic hairy roots to express *HgGLAND18i*. *HgGLAND18i*-expressing and vector control roots were inoculated with surface-sterilized *H. glycines*, and parasitic

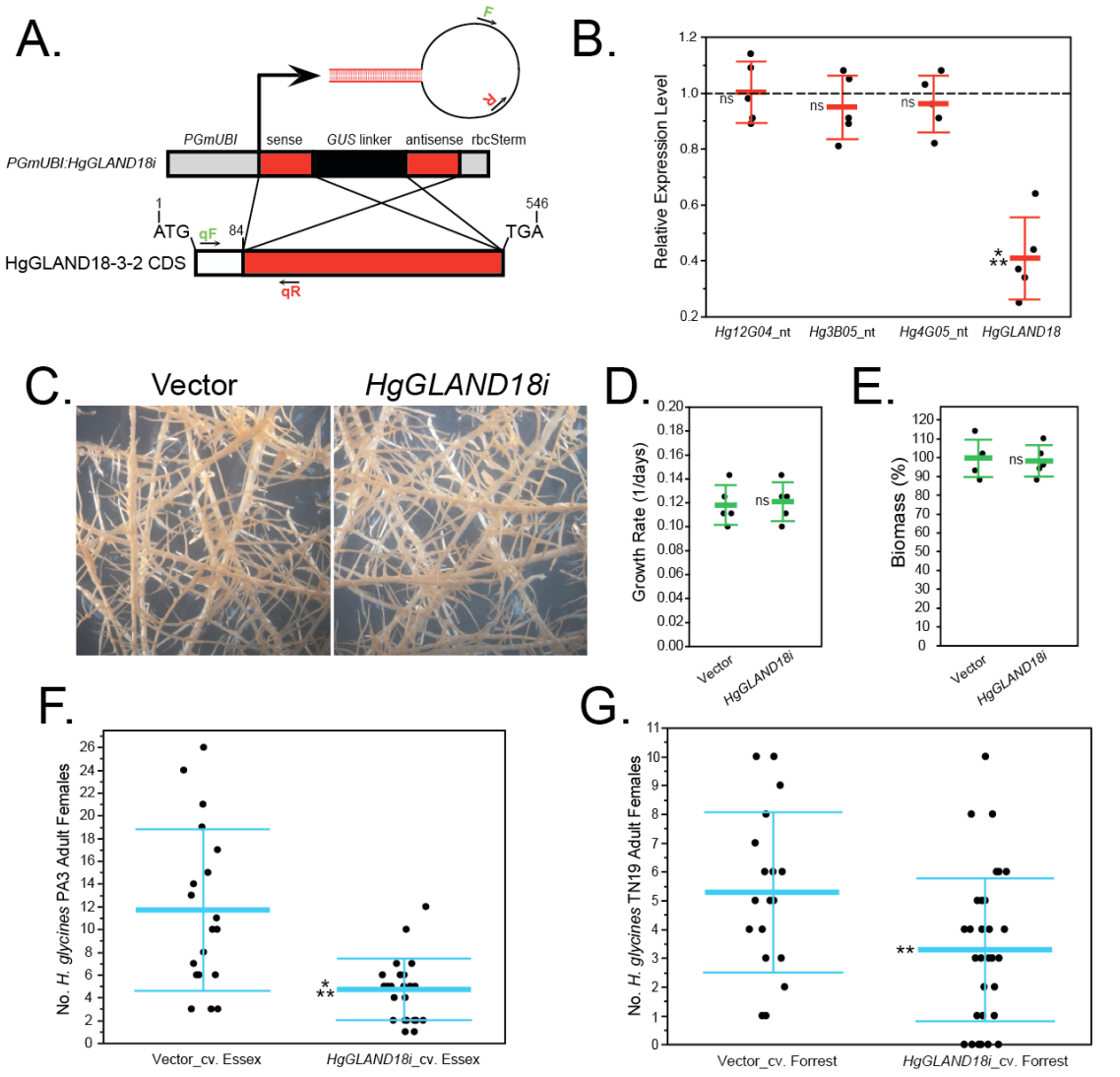



**Fig. 2. Host-induced RNA interference of *HgGLAND18*. (A)** Host-induced RNA interference (RNAi) construct generated for specifically silencing *HgGLAND18* in feeding *Heterodera glycines*. Annealing sites within the hairpin loop are shown for the primers used for diagnosis of *HgGLAND18i* transgene expression (F and R). Annealing sites for the primers used for quantitative real-time reverse-transcription (qRT)-PCR assessment of *HgGLAND18i* target gene silencing are shown (qF and qR). **(B)** qRT-PCR assessment of *HgGLAND18i* target gene silencing in *H. glycines* that fed from transgenic soybean roots. *Hg3B05* (GenBank: AF469058.1), *Hg4G06* (GenBank: AF469060.1) and *Hg8H07* (GenBank: AF500024.1) were included as non-target (nt), negative control, *H. glycines* effector genes (Sindhu et al. 2009). Expression levels of *HgGLAND18* and the non-target genes in *HgGLAND18i*-exposed *H. glycines* are relative to *H. glycines* exposed to vector control. Data were normalized to *HgActin1*. Baseline expression is set at 1.0. Five biological replicates, each representing an individual experiment, were included for all. **(C)** Qualitative and **(D,E)** quantitative growth comparisons between *HgGLAND18i*-expressing and vector control roots. (C) At least 10 independent

transgenic events were qualitatively evaluated per construct. (D,E) Data are representative of three independent experiments. (C-E) Data shown are representative of both soybean cultivars Essex and Forrest. **(F,G)** Comparisons between the number of *H. glycines* adult females that developed on *HgGLAND18i*-expressing and vector control roots. (F) Susceptible soybean cultivar (cv.) Essex inoculated with *H. glycines* avirulent line PA3 ($n = 20$). (G) Resistant soybean cv. Forrest inoculated with *H. glycines* virulent line TN19 ($n = 20$). (F,G) Data are representative of two independent experiments. (B,D-G) Data are presented as the means (thick horizontal lines) ± one standard deviation (error bars). **, $P < 0.01$; ***, $P < 0.001$; ns, not significant ($P > 0.05$).

life stages were isolated at 7-days post-inoculation (dpi). qRT-PCR detected significantly reduced *HgGLAND18* transcripts in nematodes that had infected *HgGLAND18i*-expressing roots compared to vector control roots (Fig. 2B). To test off-target effects we also analyzed the expression levels of three non-target effector genes, and none of these genes showed significant differences from vector control (Fig. 2B). Thus, in our assay, host-induced RNAi of *HgGLAND18* was successful at specifically reducing *HgGLAND18* transcripts.

We performed susceptibility assays using two different soybean cultivars (Essex or Forrest) as well as two different nematode lines (PA3 or TN19) in order to account for varietal and virulence differences in the soybean and nematode, respectively. In these experiments, *HgGLAND18i*-expressing and vector control roots exhibited similar appearance (Fig. 2C), indistinguishable growth rates (Fig. 2D) and biomass (Fig. 2E). *HgGLAND18i*-expressing soybean roots resulted in highly significant reductions in the number of *H. glycines* adult females compared to vector control regardless of soybean cultivar or nematode line (Fig. 2F,G). Taken together, these results reveal an important virulence function of HgGLAND18.

### 5.3.3 HgGLAND18 causes severe growth defects in soybean roots

To further assess the importance of *HgGLAND18* for *H. glycines* virulence, we constitutively expressed the *HgGLAND18* (isoform 3-2) CDS without the signal peptide (*HgGLAND18*$^{-sp}$) in soybean hairy roots under the *GmUBI* promoter (Fig. 3A). We did not include the signal peptide since it is most likely removed from HgGLAND18 before delivery into the plant. This manipulation resulted in severe qualitative and quantitative growth differences. Compared to the vector control, *HgGLAND18*$^{-sp}$-expressing roots grew significantly slower (Fig. 3B), generated significantly less biomass (Fig. 3C), and overall showed a *STUMPY/GLOSSY* phenotype (Fig. 3D). Because of these severe growth defects, we were unable to reliably assay these roots for changes in susceptibility to *H. glycines*.

### 5.3.4 HgGLAND18$^{-sp}$ suppresses PTI and ETI

*HgGLAND18*$^{-sp}$ was translationally fused with the type III secretion system (T3SS) signal from the AvrRPS4 effector of the *Pseudomonas syringae* pathovar *tomato* (*Pst*) strain DC3000 plant pathogen (Fig. 4A). This construct allowed the secretion of HgGLAND18$^{-sp}$ from *Pseudomonas* bacteria into colonized plant tissues and cells via the T3SS (Fabro et al. 2011). The plasmid vector was conjugated into non-pathogenic *Pseudomonas fluorescens* strain EtHAn and *Pst* DC3000 for PTI and ETI suppression experiments, respectively. Prior to inoculation, the bacteria were grown in T3SS-inducing medium, pelleted, and the supernatants were confirmed to contain HgGLAND18$^{-sp}$, while a strong common band in the pellets was not detected in the supernatants (Fig. 4B). These

preliminary control analyses indicated the secretion of HgGLAND18$^{-sp}$ from both bacteria via the T3SS (Fabro et al. 2011).

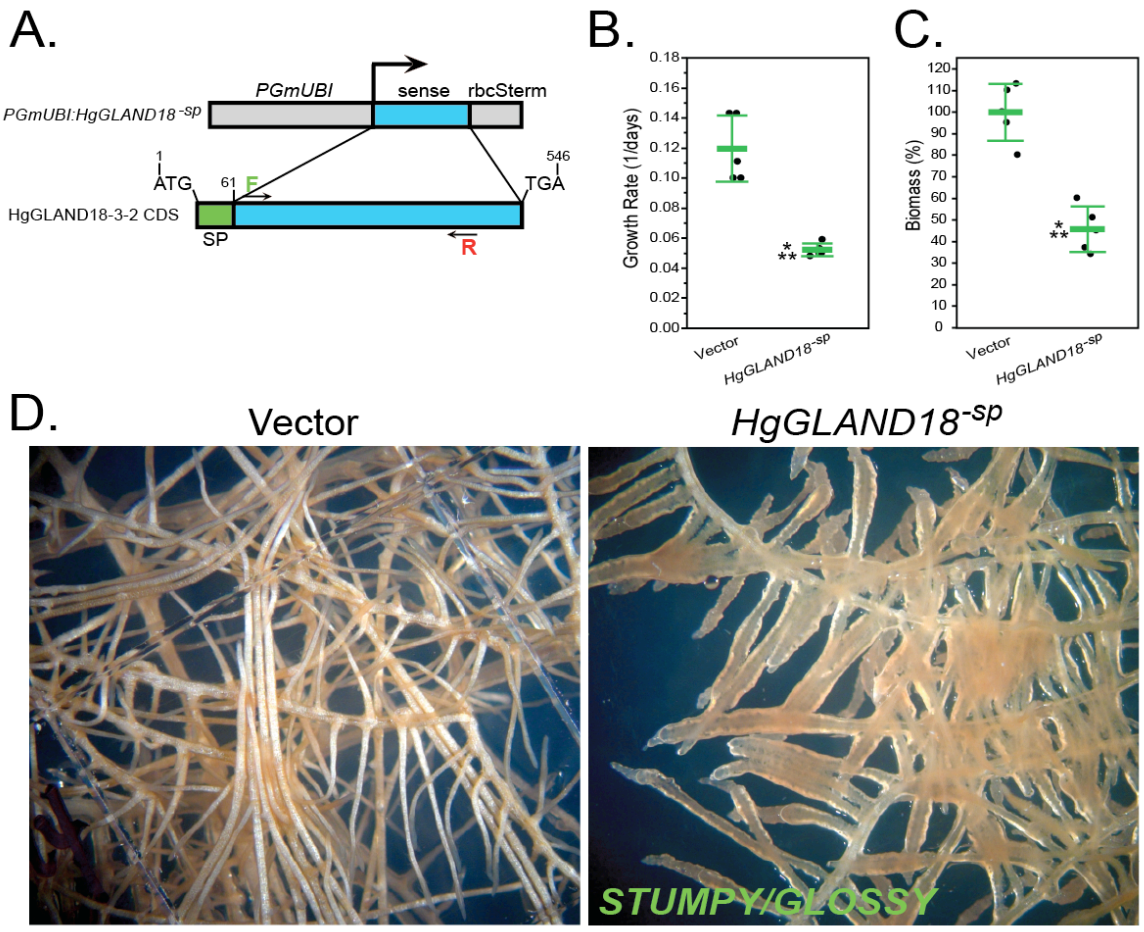



**Fig. 3. Ectopic expression of *HgGLAND18* in soybean roots. (A)** Construct generated for ectopic expression of *HgGLAND18* minus signal peptide (*HgGLAND18*$^{-sp}$) in soybean roots. Annealing sites for the primers used for diagnosis of *HgGLAND18*$^{-sp}$ transgene expression are shown (F and R). **(B,C)** Quantitative and **(D)** qualitative growth comparisons between *HgGLAND18*$^{-sp}$-expressing and vector control roots, as in Figure 2C-E.

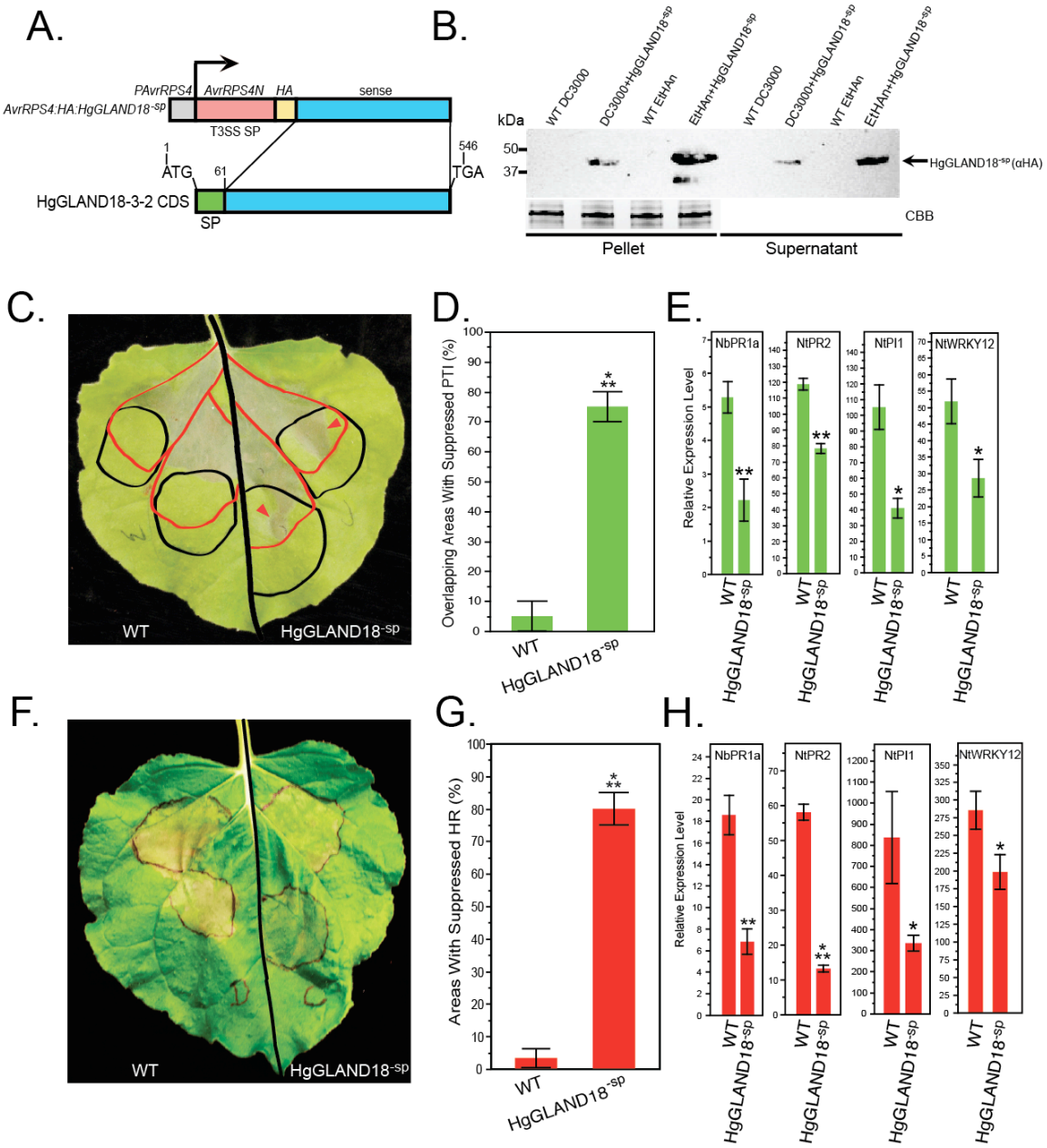



**Fig. 4. HgGLAND18 suppresses both pattern- and effector-triggered immunity. (A)** Construct generated for HgGLAND18 minus signal peptide (HgGLAND18$^{-sp}$) expression in and secretion from *Pseudomonas* into *Nicotiana benthamiana* for pattern- and effector-triggered immunity (PTI and ETI, respectively) suppression experiments. **(B)** Western blot showing specific expression of HgGLAND18$^{-sp}$ in (pellet) and secretion from (supernatant) both *Pseudomonas syringae* pathovar *tomato* strain DC3000 (*Pst* DC3000) and *Pseudomonas fluorescens* strain EtHAn for ETI and PTI suppression experiments, respectively. Bacteria were cultured in *hrp*-inducing (type III secretion system; T3SS) minimal medium beforehand. Anti (α)-HA antibody was used for the Western blot and a strong common band present in all pellet samples from Coomassie Brilliant Blue (CBB)-stained gels was used as loading control, and this strong common band was not detected

in the supernatant. **(C,D)** PTI suppression experiments. (C) Wild-type (WT) EtHAn and EtHAn + HgGLAND18$^{-sp}$ (both OD600 = 0.2) were infiltrated into *N. benthamiana* leaves (black tracing) on opposite sides of the midrib, and after 6-hrs, challenge infiltrations were performed with WT *Pst* DC3000 (OD600 = 0.02) (red tracing). Red arrows show hypersensitive cell death response (HR) caused by *Pst* DC3000 after 2-days post-infiltration (dpi) within the overlapping areas for EtHAn + HgGLAND18$^{-sp}$, indicating a suppressed PTI response against EtHAn. (D) Comparison between the percentage of overlapping areas ($n$ = 20) with suppressed PTI (presence of HR caused by *Pst* DC3000) for WT EtHAn and EtHAn + HgGLAND18$^{-sp}$. Data were pooled from three independent experiments. **(E)** Quantitative real-time reverse-transcription (qRT)-PCR assessment of the induction of salicylic acid (SA)-responsive defense marker gene expression during PTI responses for both WT EtHAn and EtHAn + HgGLAND18$^{-sp}$ at 6-hrs post-infiltration (hpi). Expression levels are relative to mock-infiltrated leaves, and normalized to *NbActin1*. Three biological replicates were included for all, each representing an individual experiment. **(F,G)** ETI suppression experiments. (F) WT *Pst* DC3000 and *Pst* DC3000 + HgGLAND18$^{-sp}$ (both OD600 = 0.02) were infiltrated into *N. benthamiana* leaves on opposite sides of the midrib, and images were taken at 3-dpi. (G) Comparison between the percentage of infiltrated areas ($n$ = 20) with suppressed HR for WT *Pst* DC3000 and *Pst* DC3000 + HgGLAND18$^{-sp}$. Data were pooled from three independent experiments. **(H)** qRT-PCR assessment of the induction of SA-responsive defense marker gene expression during ETI responses for both WT *Pst* DC3000 and *Pst* DC3000 + HgGLAND18$^{-sp}$ at 16-hpi, as in panel (E). (D,E,G,H) Data are presented as in Figure 2. *, $P < 0.05$.

For PTI suppression assays, wild-type (WT) EtHAn or EtHAn + HgGLAND18$^{-sp}$ were infiltrated into *Nicotiana benthamiana* leaves, and infiltrated sectors then were challenged with *Pst* DC3000 (Chakravarthy et al. 2009) (Fig. 4C), which triggers HR after successful colonization. WT EtHAn-triggered PTI in infiltration zones on all leaves completely prevented the colonization by *Pst* DC3000 (no HR), while outside of the WT EtHAn zones *Pst* DC3000 caused strong HR (Fig. 4C). However, nearly all EtHAn + HgGLAND18$^{-sp}$ zones allowed the spread of HR caused by *Pst* DC3000 (Fig. 4C), which indicated suppression of PTI by HgGLAND18$^{-sp}$. These differences in the PTI response were determined to be highly significant (Fig. 4D).

In separate experiments, WT EtHAn, EtHAn + HgGLAND18$^{-sp}$, or buffer control, were infiltrated into *N. benthamiana* leaves. At 6 hours post-infiltration (hpi), we

quantified the transcripts of four SA-responsive defense marker genes via qRT-PCR. These four marker genes were *pathogenesis-related 1a* (*PR1a*), *PR2*, *WRKY transcription factor 12* (*WRKY12*) and *proteinase inhibitor 1* (*PI1*) (Liu et al. 2013). We chose 6-hpi because in pilot assays this time point was determined to be the optimum for the experiments (Fig. S3, Appendix D). All four marker genes showed significant downregulation of mRNA abundance in EtHAn + HgGLAND18$^{-sp}$ compared to WT EtHAn (Fig. 4E). EtHAn + HgGLAND18$^{-sp}$ showed increases in transcript abundances for all four marker genes compared to buffer control (Fig. 4E). Thus, PTI was initiated in EtHAn + HgGLAND18$^{-sp}$, but the magnitude of the response was significantly reduced compared to WT EtHAn.

To test the ability of HgGLAND18$^{-sp}$ to suppress ETI, WT *Pst* DC3000 and *Pst* DC3000 + HgGLAND18$^{-sp}$ were infiltrated into *N. benthamiana* leaves (Fig. 4F). After 2 and 3-dpi, *Pst* DC3000 + HgGLAND18$^{-sp}$ infiltrated zones showed suppressed HR compared to WT *Pst* DC3000 (Fig. 4F). These differences were determined to be highly significant (Fig. 4G). In separate experiments, quantification of the expression levels of the four SA-responsive defense marker genes revealed significant downregulation in the *Pst* DC3000 + HgGLAND18$^{-sp}$ infiltrated leaves compared to the leaves infiltrated with WT *Pst* DC3000 (Fig. 4H). Also, similar to the PTI experiments, comparison of the transcript levels of the marker genes for *Pst* DC3000 + HgGLAND18$^{-sp}$ with buffer control indicated that the ETI response still occurred, but to a much lower extent relative to WT *Pst* DC3000. Collectively, these results indicated that HgGLAND18$^{-sp}$ suppresses the induction of both PTI and ETI responses.

### 5.3.5 Multiple protein domains in HgGLAND18 coordinate for immunosuppression

HgGLAND18 contains an internal 43-aa stretch (aa 91-133) of mostly charged aa, which we termed supercharged domain (Fig. 5A). Because of the unique aa composition in this domain, we deleted this domain (HgGLAND18$^{-sp_\Delta 91-133}$), and both this deletion mutant and various regions of HgGLAND18 were tested for ETI suppression.

HgGLAND18$^{-sp_\Delta 91-133}$ no longer suppressed ETI, while HgGLAND18$^{91-133}$ was still active, but significantly less so than WT HgGLAND18$^{-sp}$ (Fig. 5B). We also tested constructs HgGLAND18$^{21-91}$, HgGLAND18$^{91-182}$, and HgGLAND18$^{133-182}$, none of which suppressed ETI (Fig. 5B). However, HgGLAND18$^{21-133}$ still suppressed ETI at a level between WT HgGLAND18$^{-sp}$ and HgGLAND18$^{91-133}$ (Fig. 5B). We then generated transgenic soybean hairy roots for all HgGLAND18 constructs described above, and only HgGLAND18$^{21-133}$ and HgGLAND18$^{91-133}$ phenocopied the *STUMPY/GLOSSY* phenotype observed for WT HgGLAND18$^{-sp}$ (Fig. 5C). Thus, the 70 N-terminal aa and the supercharged domain are necessary for immunosuppression, the supercharged domain alone is partially sufficient, and the 70-aa N-terminal and 49-aa C-terminal domains coordinate with the supercharged domain for the most potent effect. Also, there is an evident correlation between HgGLAND18 immunosuppression and its *STUMPY/GLOSSY* phenotype in soybean roots.

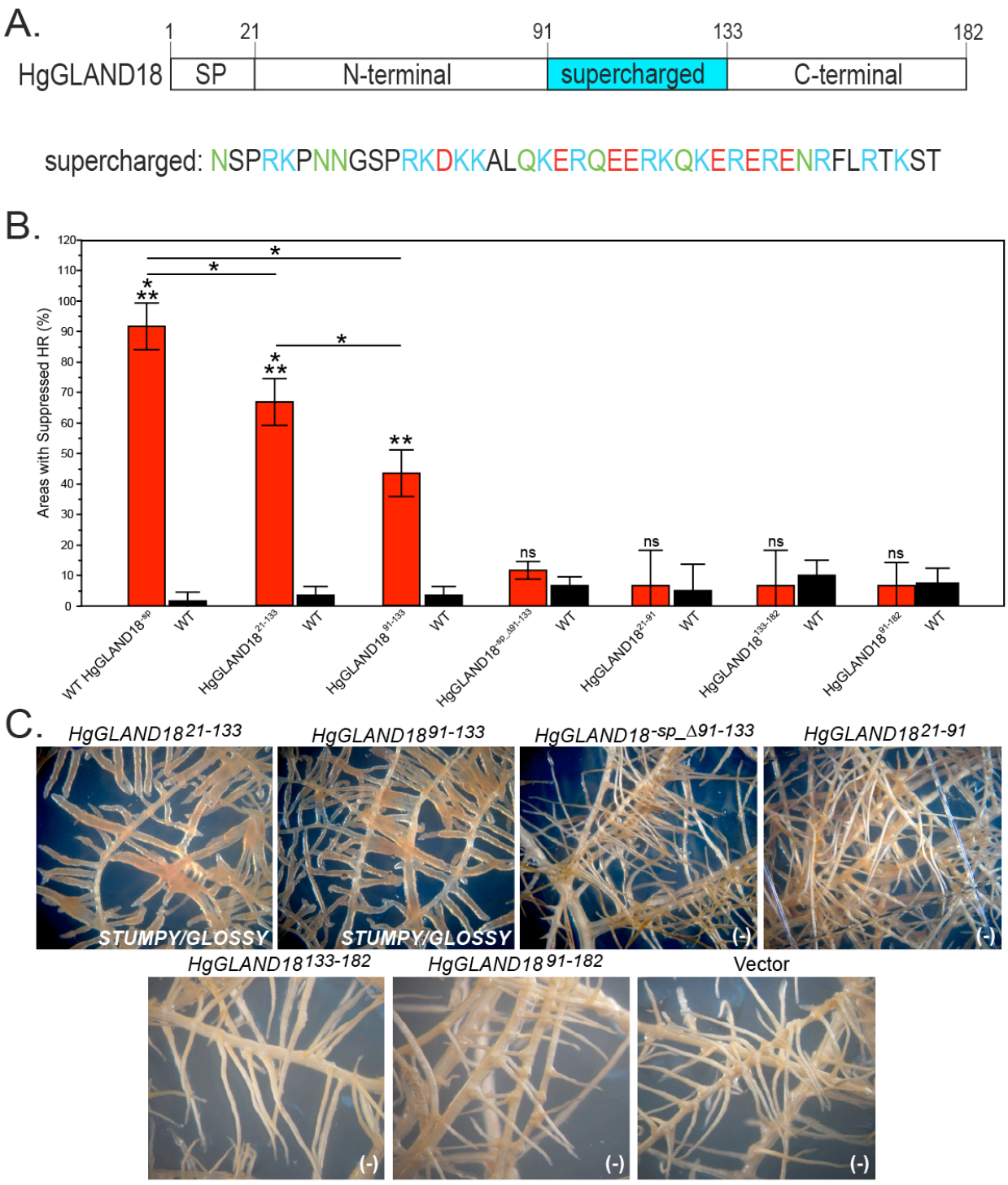



**Fig. 5. Analyses of HgGLAND18 deletion mutants. (A)** Illustration of the amino acid (aa) positions within HgGLAND18 where the supercharged domain is located. The aa sequence of supercharged is provided below the illustration with cationic and anionic aa colored light blue and red, respectively, and polar aa colored green. **(B)** Effector-triggered immunity (ETI) suppression experiments for HgGLAND18 deletion mutants,

performed as in Figure 4F,G. In addition to comparing the percentage of overlapping areas with suppressed hypersensitive cell death response (HR) between each *Pseudomonas syringae* pathovar *tomato* (*Pst*) strain DC3000 + HgGLAND18 mutant and wild-type (WT) *Pst* DC3000, comparisons were made between WT HgGLAND18$^{-sp}$ and the two HgGLAND18 mutants that also suppressed HR. Data are presented as in Figure 2. **(C)** All HgGLAND18 mutants were ectopically expressed in soybean roots as in Figure 3 for WT HgGLAND18$^{-sp}$. Images from qualitative growth comparisons are shown for all HgGLAND18 mutants and vector control roots, as in Figures 2C and 3D.

### 5.3.6 The N-terminal domain of HgGLAND18 contains marginal sequence similarity to RI, RR and RII+ domains from *Plasmodium* CSPs

The N-terminal and supercharged domains contain interesting sequence features [i.e., the former contains tandem repeats (Fig. 1B) and the latter contains mostly charged aa (Fig. 5A)], and both domains were determined to be necessary for HgGLAND18 function (Fig. 5B,C). Thus, we were next interested in determining whether other similar, but annotated sequences could be found in databases to provide putative mechanistic details. HgGLAND18 (isoform 3-5; GenBank: KT954106) was used as query in a blastp-search of nr at E-value < 0.001. This search resulted in significant similarity (E-value = 9E-12) to the *H. glycines* candidate effector HgGLAND8 (GenBank: AJR19776.1) also reported in (Noon et al. 2015). The sequence alignment covered the full-length of the sequences, but the greatest and significant alignment was within and near the signal peptides (aa 1-28).

The next highest blastp hit was a hypothetical protein from *Bacillus cereus* (GenBank: WP_000823209.1, E-value = 4E-08). In a separate blastp-search against nr using the latter as query, we identified another nearly identical *B. cereus* protein (E-value = 4E-75) named ‘circumsporozoite protein’ (GenBank: ACM13733.1), although *Bacillus* spp. do not form a sporozoite life stage. Many near identical proteins were found in other

*Bacillus* spp. Also, the similarity to HgGLAND18 was exclusive to the tandem repeats in the N-terminal domain, of which the HgGLAND18 11-aa repeat SDPI**P**I**PKQEG** aligned with the *Bacillus* protein 11-aa repeat HADL**P**A**PKQEG**. Interestingly, the blastp-searches with the *B. cereus* 'circumsporozoite protein' also resulted in significant similarity to actual CSPs from *Plasmodium simiovale*, *P. fieldi* and a *P. vivax*-like species (Table S1, Appendix D) (E-value = 5E-09, 7E-09 and 2E-08, respectively). The *B. cereus* protein repeat aligned with the tandem 11-aa repeat AAA/V**P**GAN**QEG** in the three *Plasmodium* CSPs.

Intriguingly, sequence alignments with manual inspection resulted in alignment between the HgGLAND18 N-terminal domain and the *Plasmodium* CSPs also outside of the repeats. The RI domain from *Plasmodium* CSPs aligned with the HgGLAND18 domain immediately N-terminal to the tandem repeats with 36% identity and 71% similarity (Fig. 6A,B). The RR domain from *Plasmodium* CSPs shared 36% identity and 64% similarity with the HgGLAND18 tandem repeats (Fig. 6A,B). Finally, an internal region (31-aa) of RII+ from *Plasmodium* CSPs aligned with 35% identity and 58% similarity with the HgGLAND18 domain immediately C-terminal to the tandem repeats (Fig. 6A,B). However, PEXEL/VTS, RIII and GPI-anchor domains, which have been shown to function in *Plasmodium*-specific infection processes, did not align with HgGLAND18 (Fig. 6A,B). Thus, the N-terminal domain of HgGLAND18 contains sequence similarities exclusively to RI, RR and RII+ domains from these specific *Plasmodium* CSPs.

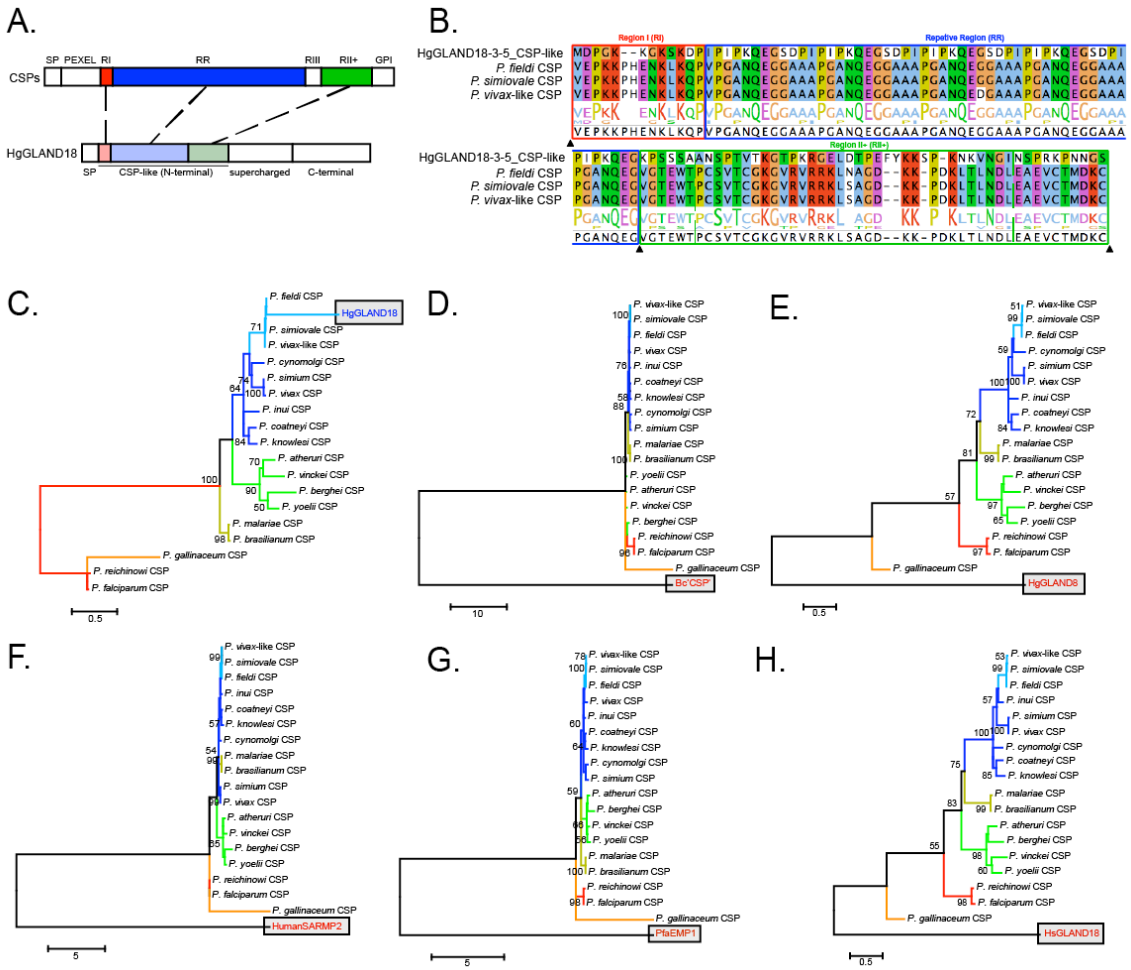



**Fig. 6. The HgGLAND18 N-terminal domain is similar to domains RI, RR and RII+ from specific *Plasmodium* CSPs. (A)** Illustration showing specific similarity of domains RI (region I), RR (repetitive region) and RII+ (region II+) from *Plasmodium* CSPs (circumsporozoite proteins) with the HgGLAND18 N-terminal (CSP-like) domain. **(B)** Multiple sequence alignment (MSA) between the HgGLAND18 N-terminal (CSP-like) domain and domains RI, RR (i.e., 5 repeats) and RII+ from *Plasmodium fieldi*, *P. simiovale* and *P. vivax*-like CSPs. Black triangles indicate the removal of the corresponding domains from the CSPs in order to generate the MSA. A consensus sequence is provided below the MSA only to indicate the identical amino acids. **(C-H)** Maximum likelihood (ML) phylogenetic trees of all eighteen *Plasmodium* CSP RI, RR and RII+ domains reported in GenBank (Table S1) with (C) HgGLAND18 [i.e., N-terminal (CSP-like) domain], (D) *Bacillus cereus* 'circumsporozoite protein' (Bc'CSP'), (E) HgGLAND8, (F) Human SARMP2, (G) *Plasmodium falciparum* EMP1, and (H) *Heterodera schachtii* GLAND18 (HsGLAND18). (D-G) Negative controls for the analysis (Materials and Methods). Bootstrap values indicate the percentage of trees ($n$ = 100) at the corresponding nodes that resulted in the same topology. Bootstrap values < 50 were removed. Scale bars indicate the rates of amino acid substitution per site. Branches for the five major *Plasmodium* clades are color coordinated as follows: *P. reichinowi*/*P.*

*falciparum* malaria clade, red; Avian malaria clade (*P. gallinaceum*), orange; African Primate malaria clade, mustard; Reptile malaria clade, light green; Asian Primate malaria clade, blue; monophyletic group of Asian Primate malarias *P. fieldi*, *P. simiovale* and *P. vivax*-like, light blue (Mitsui et al. 2010; Pacheco et al. 2012).

### 5.3.7 The observed sequence similarity between HgGLAND18 and the *Plasmodium* CSPs is significant and unique

Extensive database searches were performed to identify any other protein sequences with similarity to RI, RR and RII+ domains. In short, we performed sensitive blast-searches of NCBI databases using CSP RR domains from eighteen *Plasmodium* species reported in GenBank (Table S1, Appendix D) and the HgGLAND18 repeats. Also, we used profile-hidden Markov models to search NCBI databases with position-specific scoring matrices generated individually for *Plasmodium* CSP RI and RII+ domains. All hits were evaluated for the similarities between HgGLAND18 and the *Plasmodium* CSP domains in question (Fig. 6A,B). These searches failed to identify any sequence other than HgGLAND18 with similarity to the multiple *Plasmodium* CSP domains.

To confirm whether the similarity between HgGLAND18 and *Plasmodium* CSPs is significant (i.e., more than a random alignment), we used model selection analysis, which produces Bayesian and corrected Akaike Information Criteria (BIC and AICc) scores, to compare different models of sequence evolution by placing them into different clusters. Clustering HgGLAND18 with *Plasmodium* CSPs produced much better BIC and AICc scores than clustering HgGLAND18 with the *Bacillus* proteins mentioned above (Table S2, Appendix D). These findings indicate that HgGLAND18 is more similar to the *Plasmodium* CSPs than to the *Bacillus* proteins. In a second analysis, we tested whether

HgGLAND18 was more likely to be specifically related to the three *Plasmodium* CSPs in question or to all *Plasmodium* CSPs in general. When HgGLAND18 was clustered specifically with CSPs from *P. fieldi*, *P. simiovale* and *P. vivax*-like, our analyses produced substantially better BIC and AICc scores than clustering with any other branch in the *Plasmodium* phylogeny (Table S2, Appendix D). Also, to further assess the significance of the supported clustering of HgGLAND18 with *Plasmodium* CSPs, we tested four control sequences identified from blastp-searches (Materials and Methods). None of these controls resulted in better scores when clustered to *Plasmodium* CSPs (Table S2, Appendix D). Furthermore, we generated Maximum Likelihood (ML) phylogenetic trees for HgGLAND18 and the four control sequences separately with the eighteen *Plasmodium* CSPs. All of the controls formed outgroups to the *Plasmodium* CSPs while HgGLAND18 clustered with bootstrap support specifically to the three *Plasmodium* CSPs in question (Fig. 6C-G). These results indicated that the HgGLAND18 N-terminal domain is significantly similar to the RI, RR and RII+ domains of the three *Plasmodium* CSPs in question.

Finally, we used HgGLAND18-3-5 as query in tblastn-searches of other plant-parasitic nematode genomic and/or transcriptomic sequence databases. No sequences from plant-parasitic nematodes other than *H. glycines* were obtained with an E-value < 0.001, not even from potato cyst nematode (*Globodera* spp.) genomes or transcriptomes, or the *Heterodera avenae* transcriptome. Unfortunately, the direct sister species of *H. glycines*, the sugar beet cyst nematode *H. schachtii* (Maafi et al. 2003), was unable to be searched due to insufficient genomic and transcriptomic sequences. Southern analysis of *H. schachtii* genomic DNA resulted in hybridization of a HgGLAND18 CDS probe with

multiple intense bands for both *H. glycines* and *H. schachtii*, but not another sedentary plant-parasitic nematode, the root-knot nematode *M. incognita* (Fig. S2, Appendix D). Collectively, these findings indicated that *GLAND18* is likely present in only the *Heterodera* genus, and possibly only a few species. To further explore this observation, we cloned the *H. schachtii* GLAND18 (HsGLAND18) homolog (GenBank: KT954108) via RT-PCR. HsGLAND18 was 85% identical to HgGLAND18 (Fig. S4, Appendix D), but the similarity to the *Plasmodium* CSP domains in question was absent from HsGLAND18. Instead a number of single nucleotide polymorphisms and insertions/deletions in HsGLAND18 were evident where the domains in question aligned in HgGLAND18 (Fig. S4, Appendix D). Also, model selection analysis using HsGLAND18 did not result in better scores when clustered to *Plasmodium* CSPs (Table S2, Appendix D) and resulted as an outgroup in the ML phylogenetic tree (Fig. 6H). Thus, these results indicate that the similarity of the HgGLAND18 N-terminal domain, which we now call the CSP-like domain, with the *Plasmodium* CSPs in question likely appeared specifically in *H. glycines*, and thus, is best explained by convergent evolution.

### 5.3.8 RI, RR and RII+ domains from *Plasmodium fieldi* CSP complement the loss of the CSP-like domain from HgGLAND18

It appeared conceivable that convergence of the HgGLAND18 and *Plasmodium* CSP protein sequences could have developed due to similar immunosuppressive functions required in their requisite pathosystems. Since we had determined that the CSP-like deletion mutant HgGLAND18$^{91\text{-}182}$ is non-functional, and that the supercharged domain alone (HgGLAND18$^{91\text{-}133}$) has a weaker function compared to when CSP-like is

present (i.e., HgGLAND18$^{21\text{-}133}$), we performed functional complementation experiments by translationally fusing RI, RR and RII+ domains from *P. fieldi* CSP in-frame to the N-terminus of the HgGLAND18$^{91\text{-}182}$ and HgGLAND18$^{91\text{-}133}$ CSP-like deletion mutants. These chimeric proteins (Fig. 7A) were then tested for ETI suppression. Remarkably, these chimeric proteins fully complemented WT HgGLAND18$^{-sp}$ and HgGLAND18$^{21\text{-}133}$ (Fig. 7B). However, neither of the controls for these chimeric proteins resulted in complementation (Fig. 7B), which indicated that the complementation of the CSP-like domain in HgGLAND18 was dependent on the sequences of the *P. fieldi* CSP domains. Finally, the *P. fieldi* CSP domains alone did not suppress ETI (Fig. 7B, RI,RR,RII+), exactly as found for the CSP-like domain alone (Fig. 5B). Taken together, these results indicated that the RI, RR and RII+ domains from *P. fieldi* CSP fully complement the CSP-like domain in HgGLAND18, and thus, strongly support the conclusion of sequence convergence due to similar immunosuppressive functions.

## 5.4 Discussion

In this study, we showed that exon 2 in *HgGLAND18* from *H. glycines* line TN10 is duplicated in line OP50 (Fig. 1D). In an outbred *H. glycines* field population, we identified four different *HgGLAND18* isoforms, of which three appeared to have produced proteins that differ in the number of exon 2 repeats ranging from 0 to 5 (Fig. 1A,B). Thus, allelic variation and/or alternative splicing of repeat exons appear to generate extensive HgGLAND18 variation; the latter process has been documented for the chorismate mutase effector of plant-parasitic nematodes (Yu et al. 2011). Inter and

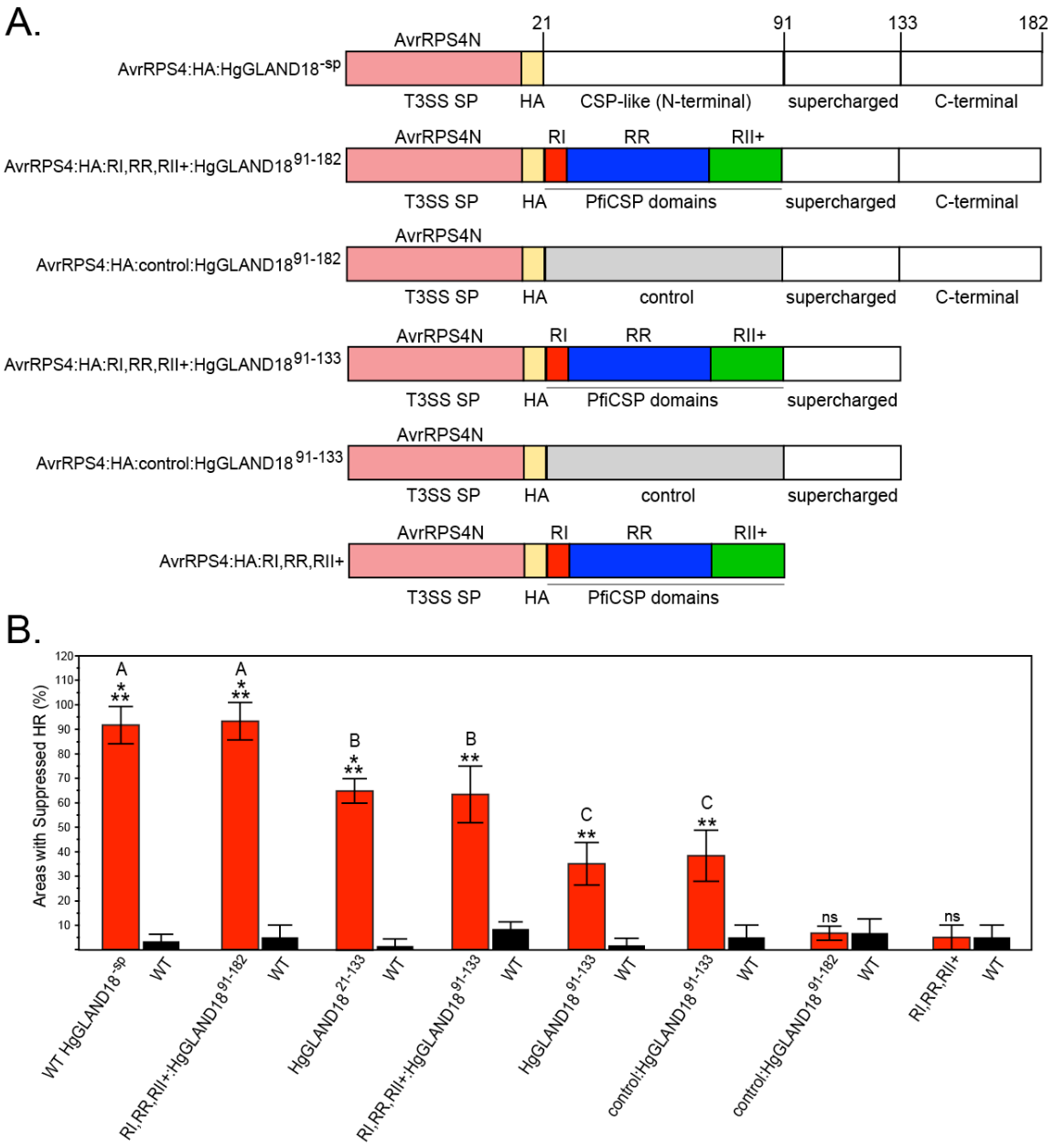



**Fig. 7. Complementation of the CSP-like domain in HgGLAND18 with domains RI, RR and RII+ from *Plasmodium fieldi* CSP. (A)** *Plasmodium fieldi* CSP (circumsporozoite protein) domains RI (region I), RR (repetitive region) and RII+ (region II+) (RI,RR,RII+) were fused and substituted in-frame for the CSP-like domain in HgGLAND18, and all chimeric proteins that were tested for complementation of immunosuppression are shown with the wild-type (WT) HgGLAND18 minus signal peptide (HgGLAND18$^{-sp}$) provided above for reference. A sequence from *GUSPlus* of the same size as the substituted RI,RR,RII+ sequence was used as a random, negative control sequence for the experiments. RI,RR,RII+ alone was also included as a negative control. **(B)** Effector-triggered immunity (ETI) suppression experiments for RI,RR,RII+ and control chimeric proteins, performed as in Figure 4F,G, with statistical cross

comparisons as in Figure 5B, but shown as significance groups. WT HgGLAND18$^{-sp}$, HgGLAND18$^{21-133}$ and HgGLAND18$^{91-133}$ were included in the experiments for comparisons. Data are presented as in Figure 2.

intra-population variation in the number of repeats has been documented for other cyst nematode effectors (Eves-van den Akker et al. 2014), and this feature may be of critical importance for infection. Importantly, HgGLAND18 isoform 3-2 is strongly expressed at each individual stage of the *H. glycines* life cycle, while all other isoforms are much less abundant (Fig. 1C,E). Thus, although there appears to be extensive variation in HgGLAND18, only a particular isoform(s) may be of critical importance during infection.

Multiple effectors from plant-parasitic nematodes have been shown to suppress PTI and/or ETI-related responses, and their mechanisms include scavenging reactive oxygen species (Chen et al. 2013; Lin et al. 2016), non-photochemical quenching (Lozano-Torres et al. 2014), and less well-understood mechanisms (Chronis et al. 2013; Ali et al. 2015a; Ali et al. 2015b; Chen et al. 2015). Some of these effectors can even activate immune responses (Lozano-Torres et al. 2012; Ali et al. 2015a; Ali et al. 2015b). In heterologous assays, we found that HgGLAND18$^{-sp}$ strongly suppresses both canonical PTI and ETI responses (Fig. 4). For deletion mutagenesis experiments, we only focused on ETI suppression for HgGLAND18$^{-sp}$ mutants because WT HgGLAND18$^{-sp}$ suppressed the induction of all four SA-responsive defense marker genes similarly during both PTI and ETI (Fig. 4E,H). We found that HgGLAND18 immunosuppression requires both the N-terminal CSP-like domain and the internal supercharged domain (Fig. 5A,B). The supercharged domain was also found to be partially sufficient for immunosuppression resulting in an about 2-fold less effect than WT HgGLAND18$^{-sp}$

(Fig. 5B). Addition of the CSP-like domain to the supercharged domain increased immunosuppression to a level in between supercharged alone and WT HgGLAND18$^{-sp}$ (Fig. 5B). Thus, HgGLAND18 immunosuppression requires the coordination of the CSP-like and C-terminal domains with the supercharged domain for the strongest effect. We hypothesize that HgGLAND18 suppresses both PTI and ETI by targeting a conserved point in the pathways conditioning these responses, which may not be surprising given the extent of overlap (Jones and Dangl 2006; Spoel and Dong 2012; Newman et al. 2013), and that such a function has been proposed before for the ubiquitin carboxyl extension protein effector from cyst nematodes (Chronis et al. 2013).

Consistent with an important role in infection, RNAi of *HgGLAND18* (Fig. 2A-C) decreased *H. glycines* virulence (Fig. 2F,G). For this analysis, we designed two separate experiments to scrutinize HgGLAND18 function. Since the usual *R*-gene-mediated plant pathogen resistances involve HR triggered by ETI, the two separate experiments were designed to deduce whether or not HgGLAND18 suppresses soybean resistance to *H. glycines*. In the first experiment, susceptible cultivar Essex was infected with *H. glycines* line PA3, which has no ability to overcome any known soybean resistance genes and thus is termed 'avirulent' on resistant soybean cultivars. Silencing of *HgGLAND18* in this experiment resulted in reduced *H. glycines* virulence (Fig. 2F) indicating that even in soybean–*H. glycines* interactions in which no major resistance genes have been shown to be present, *H. glycines* virulence is supported by the effector function. In the second experiment, resistant cultivar Forrest was infected with *H. glycines* line TN19, which has the ability to overcome the Forrest resistance and thus is termed 'virulent'. If HgGLAND18 is an effector conveying virulence in a specific manner to line TN19 (e.g.,

to suppress Forrest resistance), then silencing in this experiment should reduce line TN19 virulence on cultivar Forrest, but not that of line PA3 virulence on cultivar Essex. Because reduced virulence was observed in both experiments (Fig. 2F,G), we conclude that HgGLAND18 is not an effector specifically conveying virulence on resistant soybean cultivars, but is an effector that, likely, broadly suppresses immune responses in compatible interactions. It could be argued that if HgGLAND18 suppresses ETI, then it should suppress host resistance. However, soybean resistance to *H. glycines* has been demonstrated to be different than the usual *R*-gene-mediated plant pathogen resistances, involving gene networks not identified in other pathosystems (Cook et al. 2012; Liu et al. 2012). Moreover, it has been proposed that at least some plant pathogen resistances may actually be disconnected from HR triggered by ETI, and rather, be due to non-immune processes, and that suppression of HR may be important for compatible interactions (Coll et al. 2011). Thus, it is plausible that HgGLAND18 suppression of both PTI and ETI is relevant for the compatible interaction between *H. glycines* and soybean.

Ectopic expression of *HgGLAND18*$^{-sp}$ in soybean roots resulted in severe growth defects (Fig. 3). This phenotype was shown to be correlated with immunosuppression by determining that only the HgGLAND18 mutants that still suppressed immunity resulted in the same phenotype (Fig. 5C). We consider it unlikely that this phenotype was caused by overgrowth of *Agrobacterium rhizogenes* because the infected cotyledons were decontaminated in antibiotics prior to root induction, and the roots were maintained as well in media with high concentrations of antibiotics. There are tradeoffs between growth and immune responses that are generally understood to be due to limited resource availability (Huot et al. 2014). In general, growth and immune responses are inversely

related with activated immune responses suppressing growth, and vice versa (Huot et al. 2014). Thus, it can be argued that if HgGLAND18 strongly suppresses immune responses, growth should be favored. However, the overlaps between growth and immune response pathways are complex and not well understood (Huot et al. 2014). Thus, it remains possible that the observed growth defects could be a consequence of constitutive suppression of immune responses, or possibly the opposite—that the effect of HgGLAND18 on growth might cause immunosuppression. Future projects aimed at examining the transcriptional changes that occur in *HgGLAND18*$^{-sp}$-expressing soybean roots will determine the underlying causes of this phenotype.

The innate immune systems of plants and animals are mechanistically similar. Both use receptors to detect foreign invaders, and when activated, result in robust intracellular signaling to induce cellular defenses. Interestingly, the sequence and functional similarities between these plant and animal immune regulators are best explained by convergent evolution due to limited protein sequences and domains that can efficiently detect microbes in order to mount robust immune responses (Ausubel 2005; Coll et al. 2011; Maekawa et al. 2011). Here, we showed that the CSP-like domain in HgGLAND18 contains marginal sequence similarity to CSP domains RI, RR and RII+ from three closely related Asian primate malaria species (Fig. 6A,B). Also, extensive database searches did not find proteins other than HgGLAND18 that contain the extent of similarity to the multiple CSP domains. Furthermore, model selection (Table S2, Appendix D) coupled with phylogenetic analysis (Fig. 6C-H) determined that the similarity is significant and greatest to the *Plasmodium* species in question. Interestingly, the GLAND18 homolog in *H. schachtii*—the sister species of *H. glycines*—and the

paralogous effector HgGLAND8 do not contain similarity to the respective CSP domains (Fig. 6E,H; Table S2, Appendix D; Fig. S4, Appendix D). Thus, the similarity most likely appeared specifically in HgGLAND18. Moreover, the RI, RR and RII+ domains from *P. fieldi* CSP fully complemented the loss of the CSP-like domain from HgGLAND18 (Fig. 7). Collectively, our findings support a scenario whereby these effectors from highly divergent parasites of plants and animals converged on a similar protein sequence due to similar immunosuppressive functions. Thus, in addition to shaping analogous immune regulators within the immune systems of plants and animals, convergent evolution might be an important force causing even very different pathogens that infect these eukaryotes to utilize similar, but analogous effectors.

In summary, we have shown that *H. glycines* uses the virulence effector HgGLAND18 throughout its life cycle to suppress both PTI and ETI innate immune responses, and that the effector's mechanism might be comparable to that of the *Plasmodium* CSPs. As very few *Heterodera* effectors have been characterized, our findings help fill the gap in our understanding of how these nematodes are able to be such successful pathogens. Given the essential HgGLAND18 virulence roles, this work also exposes this effector as a possible target for novel control measures against *H. glycines*.

## 5.5 Authors' contributions

JBN performed or contributed to all experiments and analyses, data interpretation, and wrote the manuscript. MQ contributed to PTI and ETI experiments and performed protein secretion assays. DNS assisted with molecular biology manipulations, setup of

experiments, and data interpretation. UM performed database searches and computational analyses. SEVDA performed extensive searches of nematode raw sequence data, and data interpretation. TRM assisted with Nematological manipulations. DD supervised UM and assisted with data interpretation. TH and MGM helped design experiments, provided materials, and assisted with data interpretation. TH and TJB co-wrote the manuscript with JBN. TJB supervised the experimental work.

# CHAPTER 6. IMMUNOSUPPRESSION BY THE SOYBEAN CYST NEMATODE VIRULENCE EFFECTOR HGGLAND18 IS CORRELATED WITH EFFICIENCY OF LOCALIZATION TO THE NUCLEOLUS

A paper to be submitted to *Plant Cell Reports*

Jason B. Noon, Tarek Hewezi, Thomas J. Baum

## Abstract

HgGLAND18 is a strongly expressed effector protein necessary for full *Heterodera glycines* (soybean cyst nematode) virulence on soybean, and largely suppresses the plant innate immune system. The immunosuppressive function of HgGLAND18 minus signal peptide (HgGLAND18$^{-sp}$; CSP-like:supercharged:C-terminal) requires the internal supercharged and N-terminal CSP-like domains. Supercharged alone retains minimal immunosuppression, addition of CSP-like to supercharged (HgGLAND18$^{-sp}$_ΔC-terminal) increases immunosuppression, and the wild-type (WT) HgGLAND18$^{-sp}$ protein has strongest immunosuppression. Here, we investigated the subcellular localizations of WT HgGLAND18$^{-sp}$ and our previously constructed deletion mutants in *Nicotiana benthamiana* cells. We determined that WT HgGLAND18$^{-sp}$ localizes predominantly to the nucleolus. Also, entry of the small protein into the nucleus likely occurs through diffusion by the targeting of nucleolar components. In parallel to immunosuppression, we showed that the CSP-like and supercharged domains are both necessary for localization to the nucleolus. Interestingly, in addition to

the nucleolus, supercharged alone and HgGLAND18$^{-sp}$_ΔC-terminal frequently resulted in foci that resembled Cajal bodies. Inversely related to immunosuppression, the supercharged domain alone showed the presence of Cajal body-like (CBL) foci most often, HgGLAND18$^{-sp}$_ΔC-terminal showed significantly less CBL foci, while CBL foci were completely absent from nuclei targeted by WT HgGLAND18$^{-sp}$. The prevalence of CBL foci in the deletion mutants can be explained by decreased efficiency of localization to the nucleolus. Thus, the supercharged domain is partially sufficient for efficient nucleolar localization, addition of CSP-like increases efficiency, and WT HgGLAND18$^{-sp}$ has maximum efficiency. These findings indicate that the HgGLAND18 immunosuppressive function is correlated with efficiency of localization to the nucleolus.

## 6.1 Introduction

*Heterodera glycines*, the soybean cyst nematode, causes over US$1 billion in soybean yield losses each year in the USA alone (Koenning and Wrather 2010). *H. glycines* infects soybean roots and forms an elaborate feeding site in the vascular cylinder and sustains a prolonged interaction with its host to acquire nutrients (Hussey and Grundler 1998). *H. glycines* produces effector proteins in three elaborate, unicellular secretory gland cells, one dorsal and two subventral, associated with the esophagus (Mitchum et al. 2013). These effectors are delivered into plant tissues and cells and largely mediate successful parasitism (Mitchum et al. 2013). Effectors have a diversity of functions that collectively enhance *H. glycines* migration, feeding site formation, and/or

suppression of host innate immune responses [reviewed in (Goverse and Smant 2014; Hewezi 2015; Mitchum et al. 2013)].

HgGLAND18 is an effector protein produced specifically in the *H. glycines* dorsal gland (Noon et al. 2015), and homologs are present in only *Heterodera* cyst nematodes (Noon et al., manuscript submitted). *HgGLAND18* is strongly expressed throughout the entire *H. glycines* life cycle (i.e., egg to adult female), and adequate expression is essential for virulence on soybean in compatible interactions. Ectopic expression of HgGLAND18 minus signal peptide (HgGLAND18$^{-sp}$) in soybean roots causes severe growth defects. In heterologous assays, HgGLAND18$^{-sp}$ suppresses pattern- and effector-triggered immunity (PTI and ETI) (Noon et al., manuscript submitted), the plant's basal and robust hypersensitive cell death immune responses, respectively (Jones and Dangl 2006; Spoel and Dong 2012).

HgGLAND18$^{-sp}$ contains the following three distinct protein domains from N- to C-terminus: CSP-like, supercharged and C-terminal. The CSP-like domain contains significant sequence similarity to other immunosuppressive effectors called circumsporozoite proteins (CSPs) from the malaria parasites, *Plasmodium* spp. The supercharged domain contains mostly charged amino acids (aa). Both CSP-like and supercharged domains, but not C-terminal, are necessary for immunosuppression, and supercharged alone is partially sufficient retaining about 50% of the immunosuppression of wild-type (WT) HgGLAND18$^{-sp}$. Addition of CSP-like to supercharged (HgGLAND18$^{-sp}$_ΔC-terminal) increases immunosuppression to about 75% of WT HgGLAND18$^{-sp}$, while WT HgGLAND18$^{-sp}$ has strongest immunosuppression. Thus, CSP-like and C-terminal domains coordinate with supercharged to execute a highly

robust immunosuppressive function. Interestingly, ectopic expression of both supercharged alone and HgGLAND18$^{-sp}$_ΔC-terminal, but no other deletion mutants, in soybean roots phenocopies the growth defects caused by WT HgGLAND18$^{-sp}$ (Noon et al., manuscript submitted). These findings support the proposition that HgGLAND18's essential role in *H. glycines* virulence is to suppress host immune responses.

The size exclusive limit of the nuclear pore complex (NPC) has been estimated to be around 60-kilodaltons (kDa) (Wang and Brattain 2007). Many nuclear proteins are larger than the size exclusion limit of the nuclear pore, and thus, must be actively imported into the nucleus. Nuclear import receptors called importins (α and β) recognize basic nuclear localization signals (NLSs) for active import into the nucleus (Wirthmueller et al. 2013). Many pathogen effectors, including from phytonematodes (Elling et al. 2007), appear to enter nuclei through the conventional NLS/importin pathway (Wirthmueller et al. 2013). Some proteins are imported into nuclei indirectly through the conventional NLS/importin pathway by piggybacking other NLS-containing protein cargo (Wagstaff and Jans 2009). However, it has recently become clear that there are many alternatives to the conventional NLS/importin pathway such as the $Ca^{2+}$-sensor calmodulin, direct interaction with nucleoporin components of the NPC, and simply through diffusion (Wagstaff and Jans 2009).

The nucleolus is a subnuclear domain best known for the production, assembly and maturation of ribosomes (Shaw and Brown 2012). However, the nucleolus is also involved in cell cycle control, biogenesis and export of mRNAs (Brown and Shaw 2008), production of components for RNA and transcriptional gene silencing (Duan et al. 2012), and has been determined to be involved in cellular stress and innate immunity (Fuhrman

et al. 2009). Unlike the many documented pathways for import into the nucleus, how proteins are directed to subnuclear domains such as the nucleolus has received little attention. However, it is known that the nucleolus is structurally and functionally linked to subnuclear structures called Cajal bodies (CBs) (Shaw and Brown 2012). CBs are highly dynamic structures that are involved in the maturation of spliceosomal subcomplexes, small RNAs, and complexes for RNA-silencing (Shaw and Brown 2012).

In this report, we investigate the subcellular localization of WT HgGLAND18$^{-sp}$ and our previously constructed deletion mutants in *Nicotiana benthamiana* cells. We show that WT HgGLAND18$^{-sp}$ localizes predominantly within the nucleus. PSORTII predicts three NLSs in HgGLAND18, but their mutation does not prevent nuclear import of WT HgGLAND18$^{-sp}$. However, a much larger fusion protein of WT HgGLAND18$^{-sp}$ is unable to import into the nucleus, and is kept specifically in the cytoplasm. Given that WT HgGLAND18$^{-sp}$ is a small protein (17.9-kDa; ExPASy Compute pI/Mw) well below the size exclusion limit of the NPC, this suggests that WT HgGLAND18$^{-sp}$ enters the nucleus through diffusion. Interestingly, within the nucleus, WT HgGLAND18$^{-sp}$ localizes predominantly to the nucleolus. Also, identical to immunosuppression, localization of WT HgGLAND18$^{-sp}$ to the nucleolus requires both CSP-like and supercharged domains. Supercharged alone localizes to the nucleolus, but frequently mislocalizes in CB-like (CBL) foci, suggesting that WT HgGLAND18$^{-sp}$ routes to the nucleolus via CBs, and also has low efficiency of nucleolar localization. Also identical to immunosuppression, addition of CSP-like to supercharged increases the efficiency of nucleolar localization, while CBL foci are completely absent for WT HgGLAND18$^{-sp}$, and thus has highest efficiency. Therefore, these findings document a strong correlation

between the HgGLAND18 immunosuppressive function and efficiency of localization to the nucleolus.

## 6.2 Results

### 6.2.1 HgGLAND18$^{-sp}$ enters nuclei through diffusion

Two subcellular localization binary vector constructs (Chakrabarty et al. 2007) were generated by fusing HgGLAND18$^{-sp}$ to eGFP, one to the N-terminus and the other to the C-terminus (Fig. 1A). Vector constructs were transformed into *Agrobacterium tumefaciens* strain GV3101 (*At*GV3101), were infiltrated into *N. benthamiana* leaves and subjected to epifluorescence microscopy. As a result, HgGLAND18$^{-sp}$:eGFP and eGFP:HgGLAND18$^{-sp}$ both predominantly localized within the nucleus (Fig. 1C,D).

Although the HgGLAND18$^{-sp}$ protein is small (17.9-kDa; ExPASy Compute pI/Mw), it contains three PSORTII-predicted nuclear localization signals (NLSs) (Fig. 2A). Two monopartite NLSs were predicted in HgGLAND18, one at the N-terminus within the CSP-like domain (aa 23-29; NLS1) and the other in the middle of the protein within the supercharged domain (aa 101-107; NLS2) (Fig. 2A). The third predicted NLS (aa 105-121; NLS3) is bipartite and overlaps with predicted NLS2 within the supercharged domain (Fig. 2A). Thus, we initially hypothesized that the nuclear import of HgGLAND18$^{-sp}$ might be through the conventional NLS/importin pathway, regardless of being well below the size exclusion limit of the NPC. However, mutation of all cationic aa to alanines for each predicted NLS (Fig. 2A) independently or in combination did not inhibit nuclear import of eGFP:HgGLAND18$^{-sp}$ (Fig. 2B-H). On the other hand,

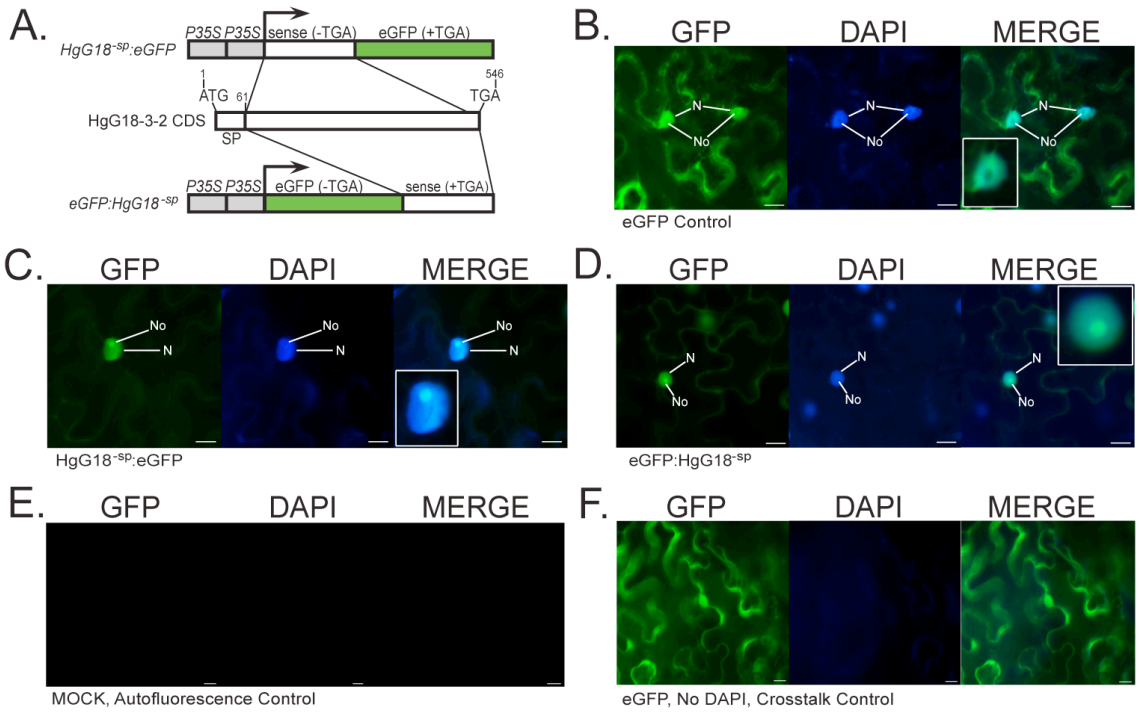



**Fig. 1. Subcellular localization of HgGLAND18$^{-sp}$ in leaf epidermal cells of *Nicotiana benthamiana*.** **(A)** Constructs generated for *in planta* subcellular localization experiments for HgGLAND18(HgG18)$^{-sp}$. **(B-F)** Subcellular localizations in *N. benthamiana* leaf epidermal cells. (B) eGFP control, (C) HgG18$^{-sp}$:eGFP, (D) eGFP:HgG18$^{-sp}$, (E) mock autofluorescence control—agroinfiltration medium alone and phosphate buffered saline w/o 4',6-diamidino-2-phenylindole (DAPI), and (F) eGFP control without DAPI staining. The latter served as a control for potential crosstalk between eGFP and DAPI channels. DAPI was used for specific staining of the nucleoplasm. Experiments were conducted with an epifluorescence microscope. Scale bars = 10-µm; N, nucleus; No, nucleolus.

each ΔNLS3 mutant did show a noticeable qualitative decrease in nuclear and increase in cytoplasmic signal (Fig. 2E-H).

Importantly, the size of the eGFP:HgGLAND18$^{-sp}$ fusion protein (46.8-kDa; ExPASy Compute pI/Mw) was still well below the exclusion limit of the NPC. We therefore generated a much larger, chimeric fusion protein containing HgGLAND18$^{-sp}$, eGFP and GUSPlus (HgGLAND18$^{-sp}$:GUSPlus:eGFP; 88.9-kDa; ExPASy Compute pI/Mw). The chimeric fusion protein was unable to import into the nucleus resulting in an

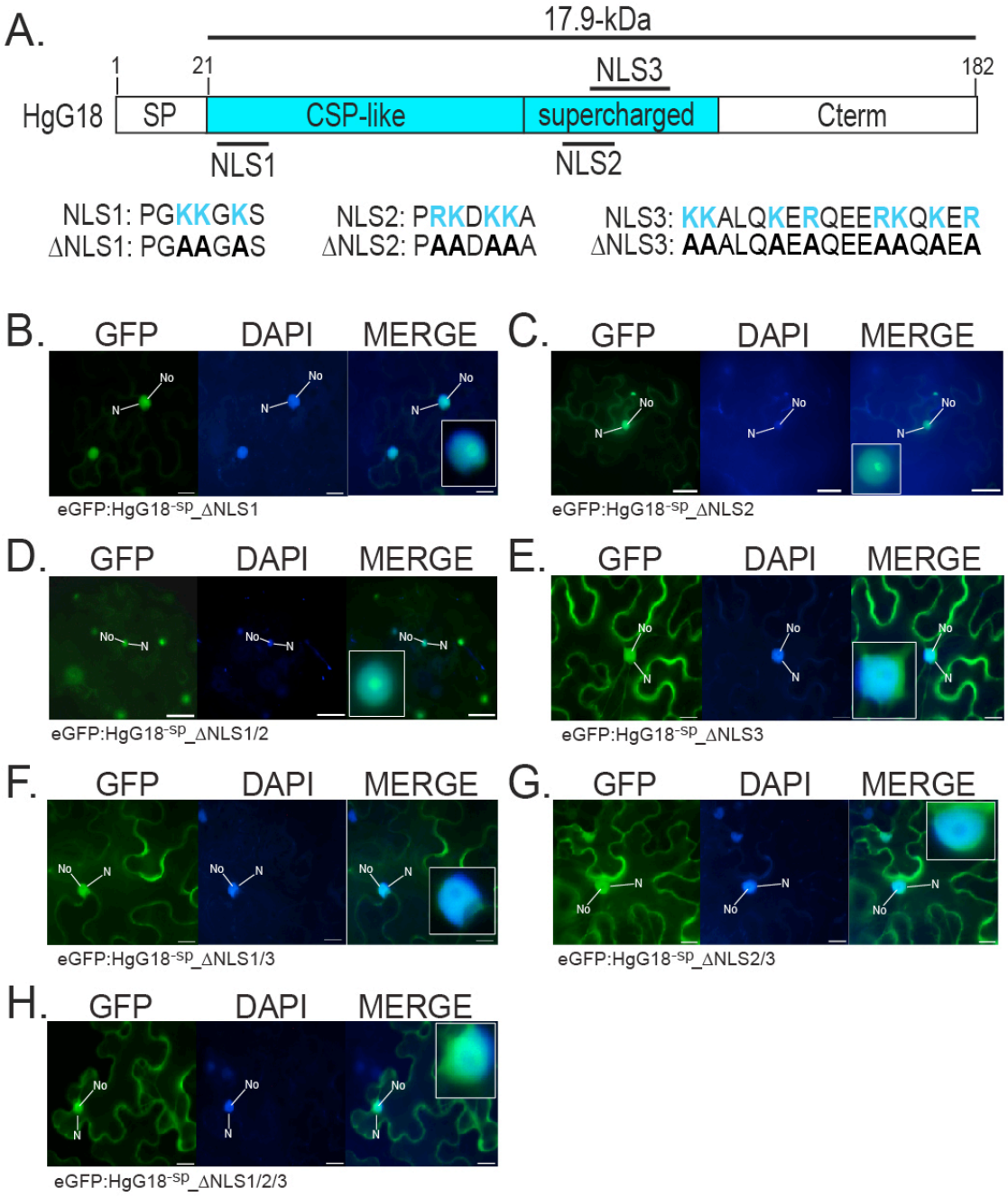



**Fig. 2. Predicted nuclear localization signals in HgGLAND18 are not required for nuclear import. (A)** Illustration of the amino acid positions within HgGLAND18 (HgG18) where predicted nuclear localization signals (NLSs) 1, 2 and 3 are located. The amino acid sequences of predicted NLSs 1, 2 and 3 are shown below the illustration, as well as the nonsynonymous mutations that were introduced into HgG18$^{-sp}$ to generate the ΔNLS1, 2 and 3 mutants. The predicted size of the HgG18$^{-sp}$ protein (17.9-kDa) is shown above the illustration. **(B-H)** Subcellular localizations of (B) HgG18$^{-sp}$_ΔNLS1, (C) HgG18$^{-sp}$_ΔNLS2, (D) HgG18$^{-sp}$_ΔNLS1/2, (E) HgG18$^{-sp}$_ΔNLS3, (F) HgG18$^{-}$

$^{sp}$_ΔNLS1/3, (G) HgG18$^{-sp}$_ΔNLS2/3, and (H) HgG18$^{-sp}$_ΔNLS1/2/3. Experiments were performed and images are presented as in Figure 1.

exclusively cytoplasmic signal (Fig. 3A) and the same was observed for the GUSPlus:eGFP negative control (Fig. 3B; 71-kDa; ExPASy Compute pI/Mw). Moreover, as a positive control, we tagged GUSPlus:eGFP with the Simian virus 40 (SV40) large T antigen NLS (SV40NLS:GUSPlus:eGFP; 71.7-kDa; ExPASy Compute pI/Mw). Addition of the SV40NLS imported the large fusion protein into the nucleus resulting in an exclusively nuclear signal (Fig. 3C). Taken together, these results indicated that the three NLSs in HgGLAND18 were falsely predicted, and that HgGLAND18$^{-sp}$ enters the nucleus most likely through diffusion. Though it can be argued that HgGLAND18$^{-sp}$ localized to nuclei simply from passive diffusion in the assay, rather than from HgGLAND18$^{-sp}$ actually targeting nuclei for its function, evidence below indicated otherwise.

### 6.2.2 CSP-like and supercharged domains are necessary for localization to the nucleolus

Within the nuclear signal for both HgGLAND18$^{-sp}$:eGFP and eGFP:HgGLAND18$^{-sp}$ was a very strong focus that was determined to be the nucleolus via absence of the 4',6-diamidino-2-phenylindole (DAPI) signal (Fig. 1C,D), which stains the nucleoplasm, but is excluded from the nucleolus (Shaw and Brown 2012). Interestingly, alanine substitutions for the cationic aa in NLS3 completely eliminated the signal in the nucleolus for mutation of predicted NLS3 alone and in all combinations with predicted NLS1 and 2 (Fig. 2E-H). Also, as mentioned above, this observation was

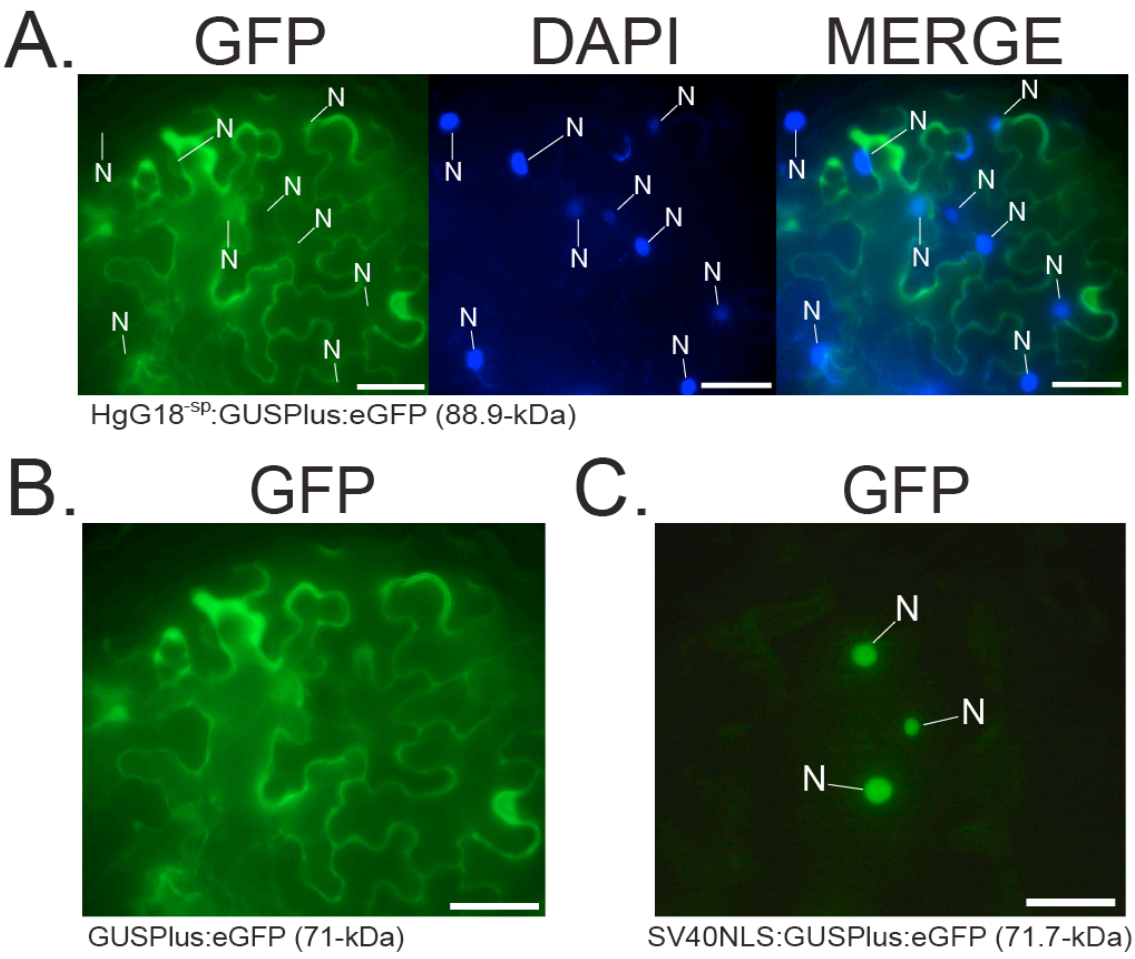



**Fig. 3. HgGLAND18$^{-sp}$ enters the nucleus through diffusion. (A-C)** Subcellular localizations of the large fusion proteins (A) HgGLAND18(HgG18)$^{-sp}$:GUSPlus (88.9-kDa), (B) GUSPlus alone (71-kDa), and (C) Simian virus 40 nuclear localization signal (SV40NLS):GUSPlus (71.7-kDa). (B,C) Negative and positive controls for the conventional NLS/importin pathway, respectively. Experiments were performed and images are presented as in Figure 1.

accompanied by a simultaneous increase in cytoplasmic signal (Fig. 2E-H). No change in nucleolar localization was observed, however, for mutation of predicted NLS1 and 2 alone or in combination (Fig. 2B-D). Thus, it appeared that the aa in the region of predicted NLS3, which is located in the supercharged domain (Fig. 2A), form a nucleolar localization signal (NoLS) that is necessary for both nucleolar localization and the predominant nuclear signal.

Since the NoLS aa in HgGLAND18 are located in the supercharged domain (Fig. 2A), we performed subcellular localization experiments on our previously generated HgGLAND18$^{-sp}$ deletion mutants (Noon et al., manuscript submitted). Fusion of the supercharged domain alone with eGFP (eGFP:supercharged) directed eGFP mostly within the nucleus with a predominant localization to the nucleolus (Fig. 4A), similar to WT HgGLAND18$^{-sp}$ (Fig. 1C,D). Conversely, deletion of the supercharged domain from HgGLAND18$^{-sp}$ (eGFP:HgGLAND18$^{-sp}$_Δsupercharged) eliminated the nucleolar signal and increased the cytoplasmic signal (Fig. 4B). The latter result was also observed, for the most part, for eGFP:CSP-like and eGFP:C-terminal (Fig. 3C,D). Noteworthy, we did notice a low percentage of nuclei (less than 5%) that showed a nucleolar signal for eGFP:CSP-like (data not shown). On the other hand, eGFP:HgGLAND18$^{-sp}$_ΔC-terminal still localized to the nucleolus (Fig. 4E), while eGFP:HgGLAND18$^{-sp}$_ΔCSP-like did not (Fig. 4F). These results indicated that *(i)* supercharged is both necessary and sufficient for localization to the nucleolus, *(ii)* addition of C-terminal to supercharged blocks nucleolar localization, and *(iii)* that CSP-like blocks the C-terminal inhibitory effect on supercharged.

### 6.2.3 Supercharged and HgGLAND18$^{-sp}$_ΔC-terminal frequently mislocalize in CBLs

Some of the nuclei targeted by eGFP:supercharged and eGFP:HgGLAND18$^{-sp}$_ΔC-terminal resulted in the presence of CBL foci in addition to nucleoli (Fig. 4A,E,G). However, CBL foci were completely absent from all nuclei that were targeted by the WT HgGLAND18$^{-sp}$ protein (Fig. 1C,D). These findings suggested that nucleolar localization

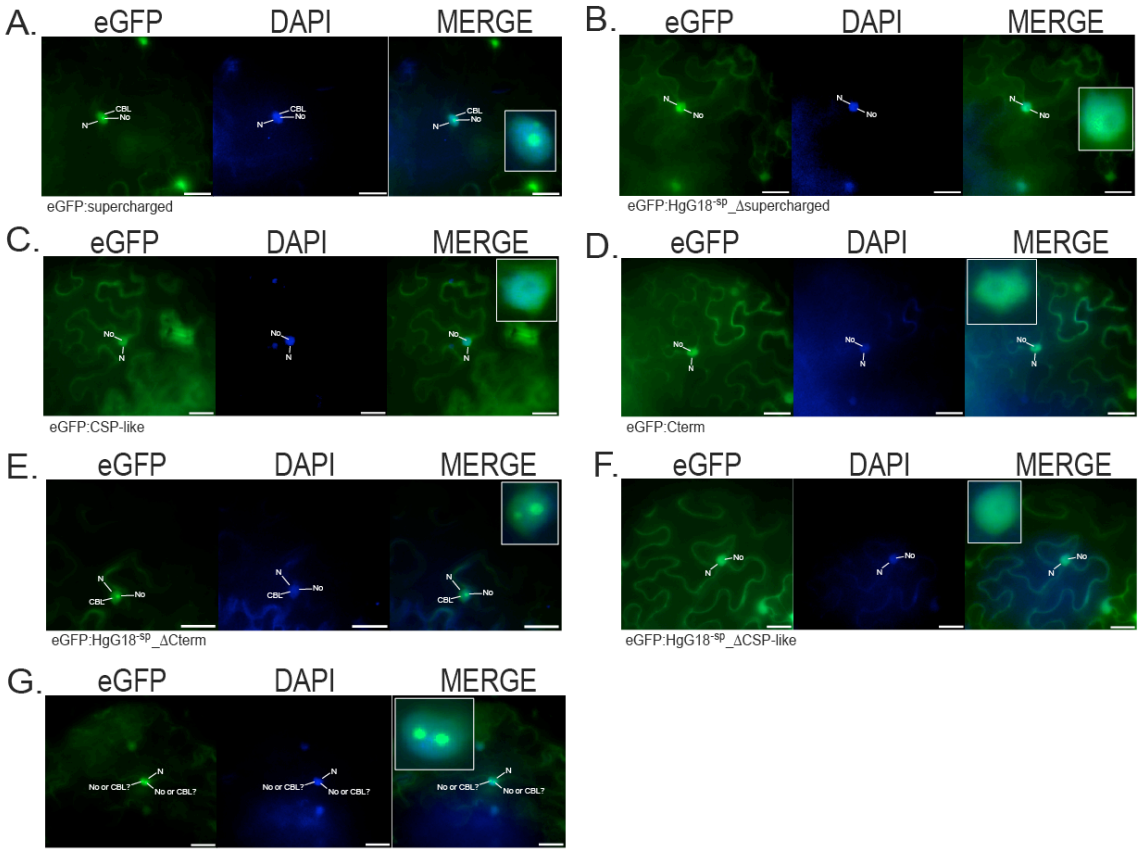



**Fig. 4. CSP-like and supercharged domains are necessary for HgGLAND18$^{-sp}$ localization to the nucleolus. (A-F)** Subcellular localizations of (A) the supercharged domain alone, (B) HgGLAND18(HgG18)$^{-sp}$_Δsupercharged, (C) CSP-like alone, (D) C-terminal(Cterm) alone, (E) HgG18$^{-sp}$_ΔCterm, and (F) HgG18$^{-sp}$_ΔCSP-like. **(G)** Representative image from subcellular localization of both supercharged alone and HgG18$^{-sp}$_ΔCterm. (A-G) Experiments were performed and images are presented as in Figure 1. CBL, Cajal body-like focus.

of HgGLAND18$^{-sp}$ involves CBs, similar to (Kim et al. 2007), and that deletion of either C-terminal or supercharged results in some nuclei where the proteins are incompletely transferred to the nucleolus. Based on these findings, we classified the presence of CBL foci in the nuclei for eGFP:supercharged and eGFP:HgGLAND18$^{-sp}$_ΔC-terminal as occurring from inefficient nucleolar localization.

### 6.2.4 Addition of CSP-like to supercharged increases efficiency of nucleolar localization and WT HgGLAND18$^{-sp}$ has maximum efficiency

Next, we calculated the percentages of nuclei that contained CBL foci for eGFP:supercharged and eGFP:HgGLAND18$^{-sp}$_ΔC-terminal. eGFP:supercharged and eGFP:HgGLAND18$^{-sp}$_ΔC-terminal resulted in averages of 30.2% and 17.7% of nuclei with CBL foci, respectively ($P < 0.05$), while eGFP:HgGLAND18$^{-sp}$ resulted in 0% (Fig. 5). Thus, although eGFP:supercharged still localized to nucleoli, its efficiency was comparatively low. When CSP-like was added to supercharged (HgGLAND18$^{-sp}$_ΔC-terminal), the efficiency of nucleolar localization was higher. Finally, when C-terminal was added, and thus WT HgGLAND18$^{-sp}$, efficiency of nucleolar localization was a maximum 100%.

### 6.2.5 Addition of CSP-like to supercharged also decreases the size of CBLs in the mislocalized nuclei

For both eGFP:supercharged and eGFP:HgGLAND18$^{-sp}$_ΔC-terminal, we frequently observed variable appearances of CBL foci (Fig. 6A). Most commonly, a single CBL focus was observed in both HgGLAND18$^{-sp}$ mutants, while two small CBL foci were very rarely observed. The appearance of the single CBL focus was either much smaller than the nucleolus or indistinguishable in size (Fig. 6A). Interestingly, we noticed that nuclei targeted by eGFP:supercharged seemed to result in more of the single large CBL focus compared to eGFP:HgGLAND18$^{-sp}$_ΔC-terminal, which seemed to contain mostly a single small CBL focus. Thus, we compared the percentages of each CBL foci appearance (i.e., one small, one large, and two small CBL foci; Fig. 6A).

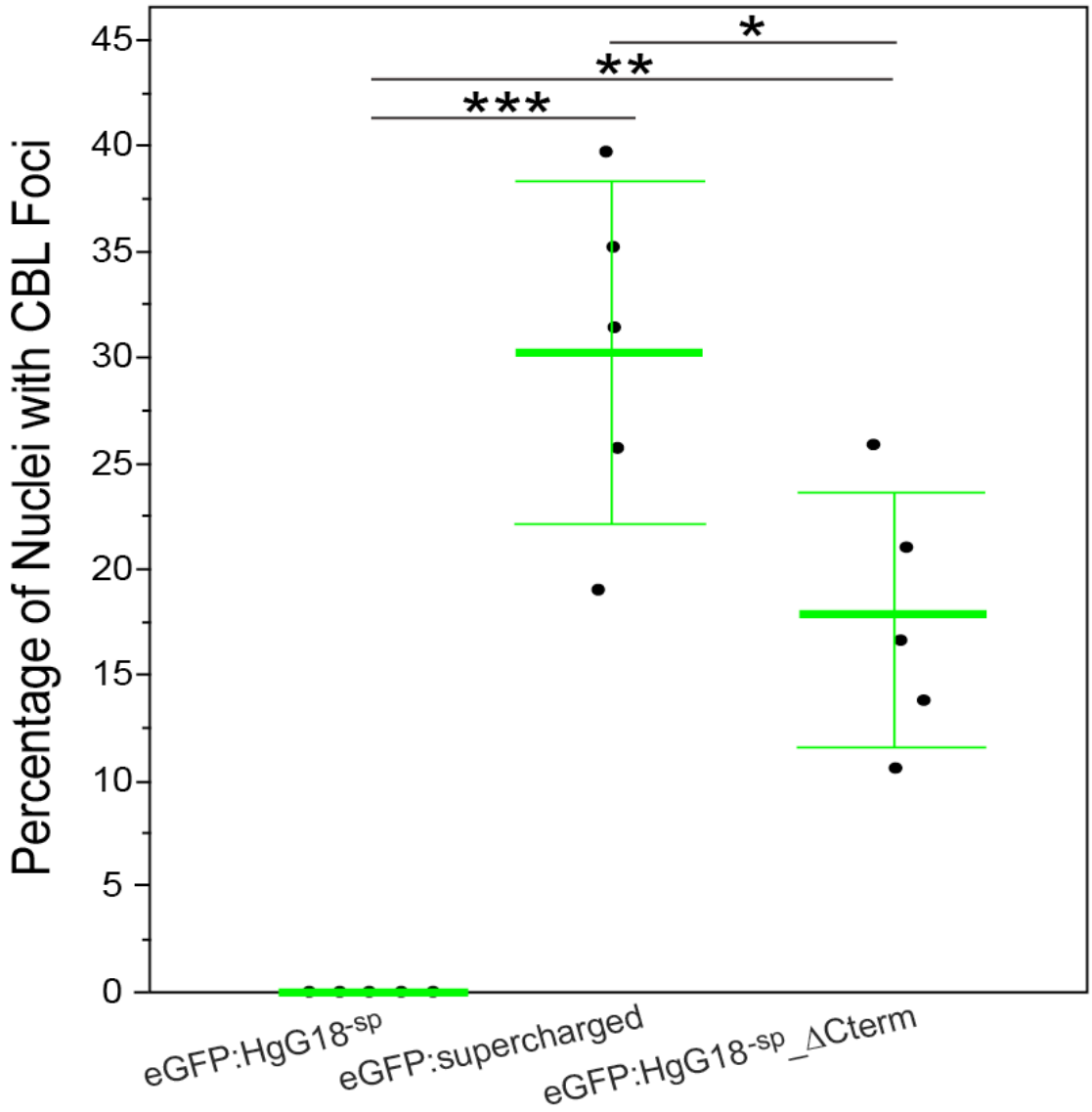



**Fig. 5. Addition of the CSP-like domain to supercharged increases efficiency of nucleolar localization.** Comparisons between the percentage of *Nicotiana benthamiana* nuclei ($n$ = 50-100) targeted by eGFP:HgGLAND18(HgG18)$^{-sp}$, eGFP:supercharged and eGFP:HgG18$^{-sp}$_ΔC-terminal(Cterm) with mislocalization in Cajal body-like (CBL) foci. Data were pooled from five independent experiments. Data are presented as the means (thick horizontal lines) ± one standard deviation (thin error bars). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

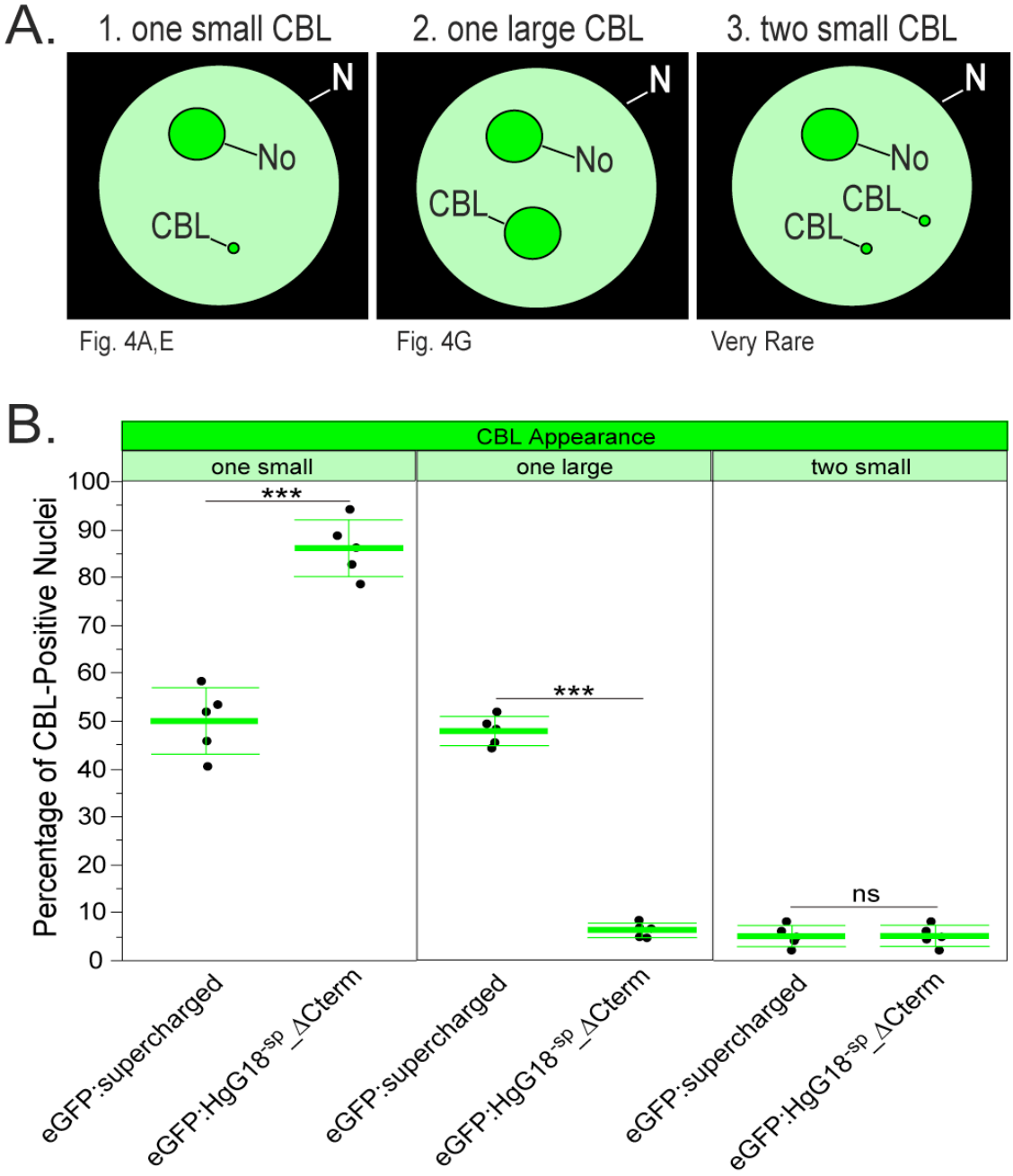



**Fig. 6. Addition of the CSP-like domain to supercharged decreases the size of CBL foci in mislocalized nuclei. (A)** Illustration of the three different Cajal body-like (CBL) foci appearances observed for the supercharged domain alone and HgGLAND18(HgG18)$^{-sp}$_ΔC-terminal(Cterm). **(B)** Comparison between the percentage of CBL-positive nuclei ($n$ = 50-100) with CBL foci that matched one of the three CBL foci appearances. Experiments were performed and data are presented as in Figure 5.

eGFP:supercharged resulted in 49.7% and 45.3% of CBL-positive nuclei with one small and one large CBL focus, respectively (Fig. 6B). However, eGFP:HgGLAND18$^{-sp}$_ΔC-terminal resulted in 88.6% and 6.4% of CBL-positive nuclei with one small and one large CBL focus, respectively (Fig. 6B). This difference in appearance of the single CBL focus between the two HgGLAND18$^{-sp}$ mutants was also highly significant (Fig. 6B, $P < 0.001$). Thus, not only did the addition of CSP-like to supercharged (i.e., HgGLAND18$^{-sp}$_ΔC-terminal) increase the efficiency of nucleolar localization, but also decreased the size of CBL foci in the mislocalized nuclei.

## 6.3 Discussion

### 6.3.1 Import into the nucleus by diffusion

In this report, we showed that HgGLAND18$^{-sp}$ is a small protein (17.9-kDa; ExPASy Compute pI/Mw), much smaller than the size exclusion limit of the NPC [~60-kDa; (Wang and Brattain 2007)]. The three predicted NLSs in HgGLAND18$^{-sp}$ are not required for nuclear import (Fig. 2), and a large HgGLAND18$^{-sp}$ fusion protein (88.9-kDa; ExPASy Compute pI/Mw) no longer imports into the nucleus, and is kept specifically in the cytoplasm (Fig. 3). Also, HgGLAND18$^{-sp}$ predominantly localizes to the nucleolus (Fig. 1) and depends on a particular signal within the supercharged domain (Fig. 4). Thus, collectively, these findings indicate that HgGLAND18$^{-sp}$ mostly likely imports into the nucleus by diffusion.

Diffusion into the nucleus becomes possible when a protein is well below the size exclusion limit of the NPC and has a target within the nucleus and/or nucleolus, such as

DNA or RNA (Lixin et al. 2001; Wagstaff and Jans 2009). Targeting of nuclear/nucleolar components favors the diffusion of the small protein into the nucleus against the concentration gradient and results in retention inside the nucleus/nucleolus (Lixin et al. 2001; Wagstaff and Jans 2009). Interestingly, the latter route into the nucleus has been thoroughly described only for the small polypeptide ligand Angiogenin (Lixin et al. 2001). Angiogenin enters the nucleus through diffusion and also predominantly localizes to the nucleolus, a process that is driven by the targeting of nucleolar components (Lixin et al. 2001; Wagstaff and Jans 2009). Thus, the localization of HgGLAND18$^{-sp}$ is remarkably similar to Angiogenin (Lixin et al. 2001; Wagstaff and Jans 2009). Also, since the RNase domain within Angiogenin is required for nucleolar localization, the targeted nucleolar components are likely RNAs (Lixin et al. 2001). Consistent with binding to nucleic acid, in addition to the identical localizations between HgGLAND18$^{-sp}$ and Angiogenin, the *in silico* protein structure of HgGLAND18$^{-sp}$ is almost completely disordered (Fig. S1, Appendix E) (Sun et al. 2013). Moreover, nucleic acid-binding prediction software such as BindN (Wang and Brown 2006) predicts HgGLAND18$^{-sp}$, mostly the CSP-like and supercharged domains, to bind RNA, but not DNA. Thus, like Angiogenin, diffusion of HgGLAND18$^{-sp}$ into the nucleus might be driven by an interaction between HgGLAND18$^{-sp}$, in particular the supercharged and possibly CSP-like domains, with nucleolar components such as RNAs.

### 6.3.2 Routing to the nucleolus via CBs

The most thorough investigation into trafficking to the nucleolus was for the ORF3 long-distance movement protein from the *Groundnut rosette virus* (Kim et al.

2007). This protein traffics to the nucleolus in *N. benthamiana* cells via CBs, causing reorganization of CBs into multiple CBL foci that fuse with the nucleolus (Kim et al. 2007). Efficient nucleolar localization is also necessary for systemic viral infection (Kim et al. 2007). We found that WT HgGLAND18$^{-sp}$ efficiently localizes to nucleoli in 100% of nuclei with no mislocalization in CBL foci (Figs. 1C,D and 5). However, although the supercharged domain alone still localizes to the nucleolus, it also mislocalizes in CBL foci in over 30% of nuclei (Figs. 4A and 5). Furthermore, addition of CSP-like to the supercharged domain decreases the occurrence of mislocalization in CBL foci to 17.5% of nuclei, and thus increases the efficiency of nucleolar localization. Consistent with ORF3 (Kim et al. 2007), these findings present a model whereby HgGLAND18$^{-sp}$ first targets CB components and then routes to the nucleolus via CBs, a process that requires the complete protein for highest efficiency. Furthermore, the order of nucleolar localization efficiency is (supercharged alone < HgGLAND18$^{-sp}$_ΔC-terminal < WT HgGLAND18$^{-sp}$), which is the same as the order of immunosuppression (Noon et al., manuscript submitted). Thus, the hierarchy of immunosuppression for HgGLAND18$^{-sp}$ and these deletion mutants is correlated with the efficiency of localization to the nucleolus.

### 6.3.3 Connection between the nucleolus and plant immunity to cyst nematodes

The ubiquitin carboxyl extension protein effector UBCEP12 from cyst nematodes, like HgGLAND18, suppresses PTI (Chen et al. 2013) and ETI (Chronis et al. 2013) and localizes in plant cell nucleoli (Tytgat et al. 2004). However, the latter studies did not mention a connection between immunosuppression and nucleolar localization. Based on

these findings, it can be suggested that HgGLAND18 and the UBCEP12 homolog in *H. glycines* might use coordinated mechanisms for immunosuppression. Conversely, the lack of a recognizable *GLAND18* homolog in the potato cyst nematodes (*Globodera* spp.) (Noon et al., manuscript submitted), which contain UBCEP12 homologs, would suggest otherwise. Yet it is possible that coordination between UBCEP12 and GLAND18 evolved specifically in *Heterodera* cyst nematodes, and such speculation is supported by the localization of *Heterodera* UBCEP12 to the nucleolus (Tytgat et al. 2004), while nucleocytoplasmic localization was shown for *Globodera* UBCEP12 (Chronis et al. 2013). Importantly, the combination of these findings for UBCEP12 and GLAND18 indicate that the nucleolus is an important compartment for innate immunity to cyst nematodes.

In the nucleolus, certain viral proteins bind host short and long double-stranded RNAs and argonaute (AGO) proteins to suppress RNA and transcriptional gene silencing (Duan et al. 2012). Also in plants, RNA silencing and associated AGO proteins are involved in innate immunity (Katiyar-Agarwal et al. 2006; Weiberg et al. 2013; Zhang et al. 2011). In the model nematode *Caenorhabditis elegans*, innate immunity to bacterial infection depends on the inhibition of resident nucleolar proteins that, in the absence of infection, sequester immune regulators to inhibit the innate immune response (Fuhrman et al. 2009). Also, certain plant immune regulators have been found to localize to nucleoli (Wang et al. 2013). Thus, these previous findings might suggest at least three possible nucleolar mechanisms for HgGLAND18's immunosuppressive function: First, ribosomes could be affected, and thus, likely global host protein synthesis; animal pathogen effectors have been shown to use such a mechanism to modulate the host immune system

(Fontana et al. 2011). Second, RNA and/or transcriptional gene silencing could be affected either by binding RNAs or protein components such as AGOs. Third, HgGLAND18 could sequester and inhibit immune regulators in nucleoli. We found that the first mechanism is unlikely as *HgGLAND18*$^{-sp}$-expressing soybean roots showed no difference in mass-normalized total protein levels compared to vector control roots (Fig. S2, Appendix E). Thus, we speculate that HgGLAND18 uses a nucleolar mechanism similar to the second and/or third suggested mechanism for immunosuppression. It is of course also possible that HgGLAND18 might be involved in cell cycle control or biogenesis and export of mRNAs (Shaw and Brown 2012). Conversely, lack of evidenced roles for these latter nucleolar processes in innate immunity might suggest otherwise. Future studies will hopefully address the exact nucleolar mechanism used by HgGLAND18 for immunosuppression, but our findings have at least provided a foundation for such studies.

## 6.4 Methods

### 6.4.1 Site-directed mutagenesis

Site-directed mutagenesis for HgGLAND18$^{-sp}$_ΔNLS mutants was performed by overlap extension PCR (Ho et al. 1989) using Platinum *Taq* (Invitrogen). As template, we used a pGEM-T Easy HgGLAND18-3-2 coding DNA sequence (CDS) clone (GenBank: KT954103; Noon et al., manuscript submitted). PCR products were cloned into pENTR using the pENTR/D-TOPO Cloning Kit (Invitrogen), and were then sequenced at Iowa State University.

### 6.4.2 Preparation of large fusion proteins

The HgGLAND18-3-2 CDS minus signal peptide (HgGLAND18$^{-sp}$) was PCR-amplified with Platinum *Taq* and subcloned into p4305.1 (GenBank: KT954098) in frame with GUSPlus. HgGLAND18$^{-sp}$:GUSPlus and GUSPlus alone were PCR-amplified with Platinum *Taq* and cloned into pENTR. SV40NLS:GUSPlus was generated by addition of SV40NLS to the forward primer (Table S1, Apendix E), and SV40NLS:GUSPlus was PCR-amplified with Platinum *Taq* and cloned into pENTR. pENTR clones were then sequenced at Iowa State University.

### 6.4.3 Vector construction

All of the above pENTR clones, and pENTR clones that we previously generated for all HgGLAND18$^{-sp}$ deletion mutants, were Gateway-cloned into pSITE-2NA (GenBank: EF212299.1) or pSITE-2CA (GenBank: EF212294.1) using LR Clonase (Invitrogen). pSITE-2NA and -2CA vector constructs were sequenced at Iowa State University, and were then transformed into *At*GV3101.

### 6.4.4 Subcellular localizations

*At*GV3101 transformed with pSITE-2NA and pSITE-2CA vector constructs were cultured overnight at 28-30˚C with shaking in 5-mL of YEP plus rifampicin (20-µg/mL), gentamicin (30-µg/mL) and spectinomycin (100-µg/mL). Overnight cultures were centrifuged at 4000-rpm for 5-min at room temperature, and pellets were resuspended in agroinfiltration medium (10-mM $MgCl_2$, 10-mM MES, pH 5.6, filter sterilized with 0.2-µm filter). Optical density at 600-nm (OD600) was measured with a Varian Cary 50 UV-

Vis Spectrophotometer, and then bacteria were adjusted to OD600s equal to 0.25. Acetosyringone was added at 200-μM (diluted in DMSO and filter-sterilized with 0.2-μm filter) to each *At*GV3101 inoculum, and then bacteria were agroinfiltrated into the abaxial surfaces of 4-5-week-old *N. benthamiana* leaves using 1-mL needless syringes (Becton, Dickinson and Company). For each construct, two leaves were infiltrated from two plants, and each leaf was infiltrated twice for a total of 8 infiltrated areas. After 48-hr, DAPI (Invitrogen) was infiltrated at 982-nM (diluted in phosphate buffered saline; PBS, pH 7.4) into the same agroinfiltrated areas using a 1-mL needless syringe in order to stain the nuclei. After 5-min, leaf disks were taken from each infiltrated area, were mounted in ddH20, and were visualized with an Axiovert 100 Epifluorescence Microscope with AxioVision Microscopy Software (Zeiss). Images were processed with Adobe Photoshop CC 2014.

### 6.4.5 Calculations of the percentage of nuclei with CBL foci

For each construct, a total of 50-100 nuclei were observed via epifluorescence in cells of leaf disks taken from 3 different *N. benthamiana* leaves in five independent experiments. Each nucleus was recorded for whether the CBL foci were present or absent in addition to the nucleolus, as well as the number (1 or 2) and size (small or large) observed (see Fig. 6A). Multiple statistical comparisons (Fig. 5) were performed by ANOVA and then Tukey-Kramer HSD post-hoc test in JMP Pro 11. Single statistical comparisons (Fig. 6B) were performed with the t-test in JMP Pro 11.

## 6.5 Authors' contributions

JBN conceived and designed research, conducted experiments, and analyzed data. TH and TJB co-wrote the manuscript with JBN. TJB supervised the work.

# CHAPTER 7. HOMEOSTASIS IN THE SOYBEAN MICRORNA396-*GRF* REGULATORY NETWORK IS REQUIRED FOR SOYBEAN CYST NEMATODE INFECTION

A paper submitted to *Plant Biotechnology Journal*

Jason B. Noon, Tarek Hewezi, Thomas J. Baum

## Summary

*Heterodera glycines*, the soybean cyst nematode, penetrates soybean roots and migrates to the vascular cylinder where it forms an elaborate feeding site called the syncytium. MicroRNAs (miRNAs) are important regulatory factors that suppress the expression of genes involved in a wide range of biological processes. miR396 targets *Growth-Regulating Transcription Factors* (*GRFs*), and the miR396-*GRF1/3* regulatory module was shown to be a master regulator of syncytium formation in the model cyst nematode interaction between the beet cyst nematode *Heterodera schachtii* and Arabidopsis. Here, we investigated whether this regulatory system operates in soybean roots and whether this system is likewise important for *H. glycines* infection. We determined that a complex network involving nine *MIR396* and twenty-three *GRF* genes is required for normal soybean root development and that *GRF* function is specified in the root apical meristem by miR396. We show that all *MIR396* genes are downregulated in the syncytium during its formation phase while eleven different *GRFs* are upregulated. Also, the switch to the syncytium maintenance phase coincides with upregulation of

*MIR396* and downregulation of *GRFs* specifically via post-transcriptional regulation by miR396. Furthermore, interference in the miR396-*GRF6/8-13/15-17/19* regulatory network, either through overexpression or knockdown experiments, prevents efficient *H. glycines* development to adult females, presumably through altered syncytia. Therefore, these results indicate that homeostasis in the miR396-*GRF6/8-13/15-17/19* regulatory network is required for efficient *H. glycines* development. Thus, this regulatory network serves as a promising target for the development of novel control measures against this devastating soybean pathogen.

## 7.1 Introduction

Cyst nematodes (*Heterodera* and *Globodera* spp.) are economically important, root-infecting, obligate biotrophs that form an elaborate feeding site within the vascular cylinder called the syncytium (Hussey and Grundler 1998; Jones et al. 2013). The syncytium serves as the single source of nourishment throughout the life of the cyst nematode. The feeding organ begins as a single cell and becomes enlarged and multinucleated through cytoplasmic fusion of numerous nearby cortical or vascular parenchyma cells via cell wall dissolution. Syncytia are characterized by dense cytoplasm, reduced vacuoles, hypertrophied nuclei and nucleoli, and abundant endoplasmic reticulum, ribosomes, plastids, and mitochondria (Sobezak and Golinowski 2009). This process of redirecting differentiated root cells into a novel developmental program ensues during a syncytium formation phase that involves immense transcriptional and post-transcriptional regulation of gene expression (Alkharouf et al.

2006; Hewezi and Baum 2012; Hewezi et al. 2012; Ithal et al. 2007b; Klink et al. 2009; Szakasits et al. 2009). The fully formed syncytium then enters a maintenance phase where no additional cells are incorporated, and thus, has completed all major developmental changes for maintaining the function of feeding the developing nematode. Interestingly, much of this reprogramming of differentiated root cells involves the concerted action of small RNAs, in particular microRNAs (miRNAs) and their target genes (Hewezi and Baum 2012; Hewezi and Baum 2015; Hewezi et al. 2008b; Hewezi et al. 2012).

In plants, miRNAs are 20-24-nt endogenous molecules that are produced from their own *MIRNA* genes and function to suppress gene expression (Rogers and Chen 2013). *MIRNA* genes are transcribed and produce a primary miRNA transcript that is first processed by DICER-LIKE 1 (DCL1) into a precursor (pre)-miRNA stem-loop structure (Bologna and Voinnet 2014). The pre-miRNA is subsequently processed by DCL1 (if a 21-nt miRNA) into short double-stranded RNAs consisting of miRNA guide and passenger (miRNA*) strands (Bologna and Voinnet 2014). The miRNA/miRNA* duplex is 2'-*O*-methylated at the 3'-ends for stabilization (Yu et al. 2005). Then, most commonly, the miRNA guide strand is loaded into the ARGONAUTE (AGO) component of the RNA-induced silencing complex (RISC) (Eamens et al. 2009). miRNA-loaded RISCs are then directed to target transcripts through miRNA/target complementarity and repress target transcripts most often through slicing or cleavage via AGO endonuclease activity (Mallory et al. 2008). miRNAs regulate the expression of transcription factors, proteins that mediate stress responses, and many other proteins that impact the development and physiology of plants (Rogers and Chen 2013).

Recent experimental evidence indicated that various miRNAs change in expression in response to cyst and root-knot nematode infection (Cabrera et al. 2016; Hewezi and Baum 2015; Hewezi et al. 2008b; Li et al. 2012; Xu et al. 2014; Zhao et al. 2015b). During infection of Arabidopsis by the beet cyst nematode *Heterodera schachtii*, many miRNAs are differentially expressed and are negatively correlated with target gene abundance (Hewezi et al. 2008b). Of these differentially expressed miRNAs, miR396 was shown to be a master regulator of syncytium formation (Hewezi et al. 2012). Also, an important role for miR390, TAS3 trans-acting short-interfering (tasi)RNAs, and their auxin response factor targets was demonstrated for root-knot nematode *Meloidogyne javanica* infection of Arabidopsis (Cabrera et al. 2016). Furthermore, the miR319-TCP4 module was shown to act as a responder and regulator of systemic defense signals, mediated by jasmonic acid, for resistance to the root-knot nematode *Meloidogyne incognita* in tomato (*Solanum lycopersicum*) (Zhao et al. 2015b). These previous findings demonstrate that miRNAs are important regulatory factors during infection by cyst and root-knot nematodes (Hewezi and Baum 2015).

Deep sequencing efforts have revealed that miRNAs in soybean (*Glycine max*) are differentially expressed during seed development, flowering time, and in the shoot apical meristem (Li et al. 2015; Shamimuzzaman and Vodkin 2012; Wong et al. 2011). Soybean miRNAs are also differentially expressed during various abiotic (Fang et al. 2013; Kulcheski et al. 2011; Wang et al. 2013; Xu et al. 2013) and biotic stress conditions. These biotic stress conditions include rot and rust diseases (Guo et al. 2011; Kulcheski et al. 2011), and interestingly, *H. glycines* infection (Li et al. 2012; Xu et al. 2014). However, the latter studies revealed only limited miRNA profiles. Collectively

though, these deep sequencing efforts suggest that miRNAs are involved in a wide range of important processes in soybean, including *H. glycines* infection. However, the only process in soybean experimentally confirmed to involve miRNAs that were initially detected through deep sequencing efforts is nodulation (Li et al. 2010a; Van et al. 2013; Yan et al. 2015).

miR396 targets the plant-specific *Growth-Regulating Factors* (*GRFs*), transcription factors characterized by the QLQ protein-interaction and WRC DNA-binding domains (Kim et al. 2003). The Arabidopsis miR396-*GRF* regulatory module is important for developmental and stress-related processes. These processes include leaf growth and development, flower development and time control, variation of seed weight and oil content, root development and physiology, longevity, and the coordination of plant growth with stress responses (Bao et al. 2014; Debernardi et al. 2014; Hewezi and Baum 2012; Hewezi et al. 2012; Horiguchi et al. 2011; Kim et al. 2012; Kim et al. 2003; Kim and Lee 2006; Liang et al. 2014; Liu et al. 2012a; Liu et al. 2014; Pajoro et al. 2014; Rodriguez et al. 2015; van der Knaap et al. 2000). But the most generalized function is the regulation of cell proliferation and expansion (Bao et al. 2014; Horiguchi et al. 2011; Kim et al. 2012; Kim et al. 2003; Liang et al. 2014; Omidbakhshfard et al. 2015; Pajoro et al. 2014; Rodriguez et al. 2015; Rodriguez et al. 2010). Homeostasis of the miR396-*GRF* regulatory module in both Arabidopsis and *Medicago truncatula* is necessary for normal root development (Bazin et al. 2013; Hewezi et al. 2012; Rodriguez et al. 2015). Interestingly, Arabidopsis *GRF1* and *3*-mediated gene expression regulation likely account for almost 50% of the genes that are differentially expressed in the *H. schachtii* syncytium (Hewezi et al. 2012; Szakasits et al. 2009), and interfering with the miR396-

*GRF1/3* regulatory module results in decreased syncytium size and arrested nematode development (Hewezi et al. 2012). Thus, the miR396-*GRF* regulatory module serves as a possible target for developing novel control measures against cyst nematodes.

In this paper, we explore the hypothesis that a miR396-*GRF* system operates in the roots of soybean and is important for *H. glycines* syncytium formation, in particular for *H. glycines* development to adult females. Using a combination of molecular and genetic analyses, we first determine that a complex network involving nine *MIR396* genes and twenty-three *GRF* genes operates in soybean roots. Interference of this regulatory network modifies root development. Also, we determine that a network involving all nine *MIR396* and eleven different *GRF* genes delineates the syncytium formation phase, which begins with a miRNA396 downregulation and a resulting *GRF* upregulation. During the switch to the syncytium maintenance phase, a miR396 expression spike in the syncytium post-transcriptionally silences *GRFs*. Furthermore, we indicate that interference in the homeostasis of this network prevents efficient *H. glycines* development to the adult female stage. Therefore, we document this network as a target for control measures against *H. glycines*.

## 7.2 Results

### 7.2.1 The *MIR396* gene family is active in young soybean roots

Eleven pre-miR396 sequences were found in the public miRNA database [miRBase (Kozomara and Griffiths-Jones 2014)] for soybean (pre-miR396a-k). These eleven pre-miR396 sequences were blastn-searched against the soybean cultivar (cv.)

Williams 82 genome sequence (Schmutz et al. 2010). Unique genomic coordinates for all but pre-miR396d and g were identified (Table S1, Appendix F). Pre-miR396d and g were found at the same genomic location, and thus, the latter was removed from our list for further analyses. As a result, ten unique *MIR396* loci most likely constitute the *MIR396* gene family in soybean cv. Williams 82.

Stem-loop structures were evaluated *in silico* for all pre-miR396 sequences in order to search for a miR396/miR396* duplex within the stems. Modeled stem-loop structures for all but pre-miR396h formed a miR396/miR396* duplex within the stems (Fig. 1A), and thus, contained a feature of a functional *MIR396* gene (Rogers and Chen 2013). Pre-miR396h, however, did not form a miR396/miR396* duplex; the 3'-end of the miR396 sequence was placed within the loop (Fig. 1A). Thus, *MIR396h* is most likely nonfunctional. Therefore, this *in silico* analysis suggested that there are nine functional *MIR396* genes in soybean cv. Williams 82.

To determine the evolutionary relationship between the different *MIR396* genes, we performed a Maximum Likelihood phylogenetic analysis. This analysis resulted in four well-supported subfamilies (Fig. 1B, bootstrap > 87). Subfamily 1 contained pre-miR396a and i, subfamily 2 contained pre-miR396e, j and h, subfamily 3 contained pre-miR396c and f, and subfamily 4 contained pre-miR396b, d and k (Fig. 1B). Pre-miR396d and k within subfamily 3 were determined to be identical, resulting in a total of eight unique pre-miR396 sequences (Fig. 1B). Also, pre-miR396 sequences within all four subfamilies were almost identical with only a few nucleotide mismatches within the loops. Furthermore, pre-miR396 within subfamilies 1 and 2 contained a miR396 sequence with a guanine at position 21 (miR396a/i/j) while subfamilies 3 and 4 contained

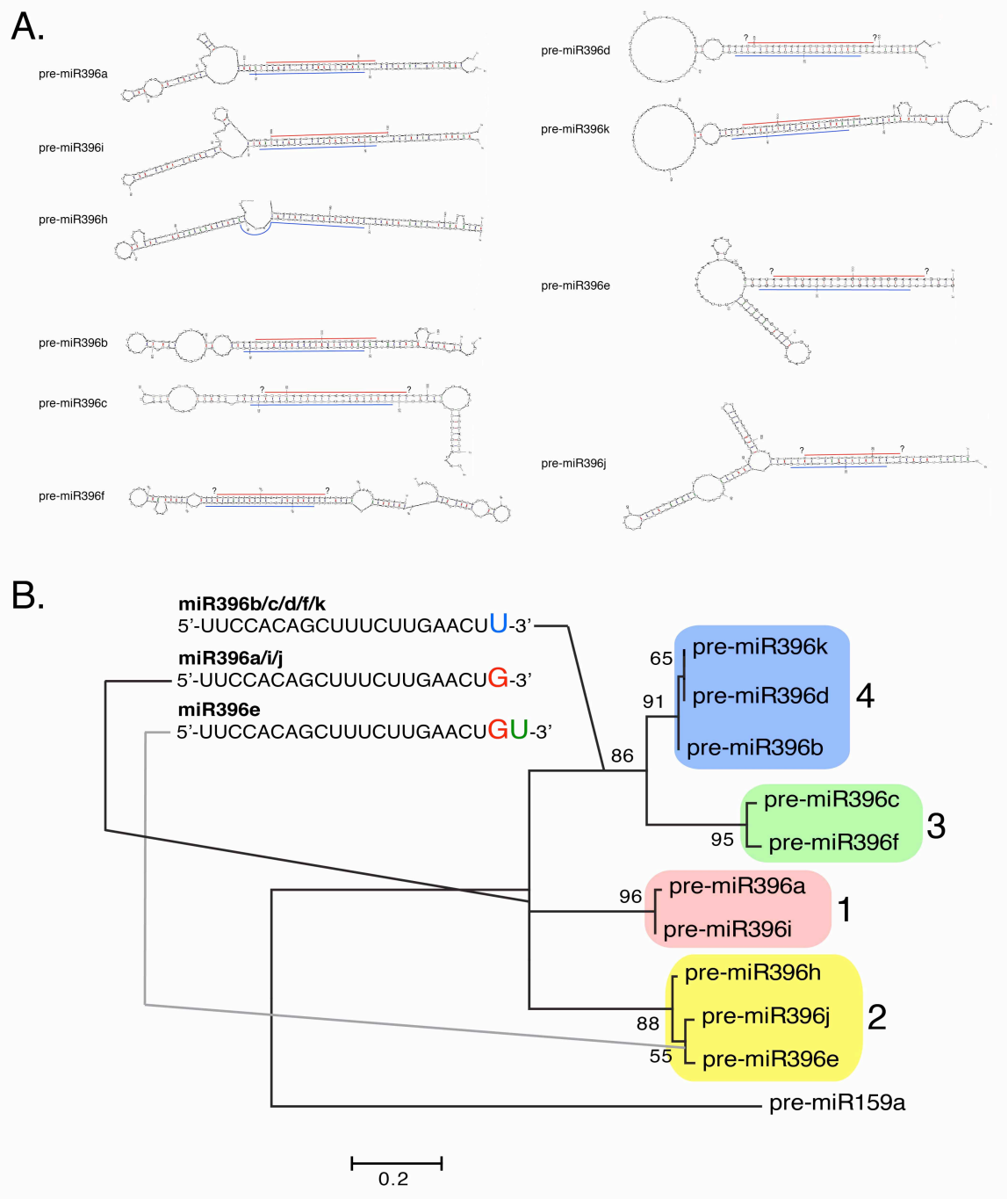



**Fig. 1. The soybean *MIR396* gene family. (A)** *In silico* pre-miR396 stem-loop structures grouped according to structural similarities. miR396 and miR396* are indicated with blue and red lines, respectively. Question marks on miR396* indicate that the exact 5'- and 3'-ends are unknown due to lack of available sequence information in miRBase. **(B)** Maximum likelihood phylogenetic tree of the soybean *MIR396* gene family. The four well-supported pre-miR396 subfamilies are indicated. The timing of appearance of each miR396 molecule is indicated at the corresponding branch. The line connecting miR396e

to the corresponding branch is colored grey to indicate a more recent appearance. Scale bar equals the number of nucleotide substitutions per site.

a uracil (miR396b/c/d/f/k) (Fig. 1B). Interestingly, pre-miR396e was the only soybean pre-miR396 found in miRBase with an additional nucleotide (nt) on the miR396a/i/j molecule producing a 22-nt miR396e molecule with a uracil at position 22 (Fig. 1B). This may suggest that pre-miR396e is processed by DCL2 instead of DCL1 and that it is loaded into a different AGO (Rogers and Chen 2013).

Next, based on the phylogenetic tree (Fig. 1B), primers were designed to target each pre-miR396 subfamily, and separate primers were designed for each miR396 molecule. Quantitative real-time reverse transcription (qRT)-PCR was performed on RNA isolated from the roots of 10-day-old soybean cv. Williams 82 seedlings. All pre-miR396 subfamilies and individual miR396 molecules resulted in detectable expression (Fig. 2A,B). Also, significant differences in expression levels were detected among both pre-miR396 and miR396. Pre-miR396a/i (subfamily 1) and pre-miR396c/f (subfamily 3) showed over 9- and 18-fold greater expression, respectively, compared to pre-miR396e/h/j (subfamily 2) and pre-miR396b/d/k (subfamily 4) (Fig. 2A). miR396a/i/j and miR396b/c/d/f/k molecules showed 4- and 12-fold greater expression compared to miR396e (Fig. 2B). Taken together, these results indicated that the *MIR396* gene family was active with variable levels of expression in young, developing soybean roots.

### 7.2.2 Twenty-three *GRFs* contain putative miR396 target sites and are active in young soybean roots

The *GRF* gene family has been described for soybean (Omidbakhshfard et al. 2015). To

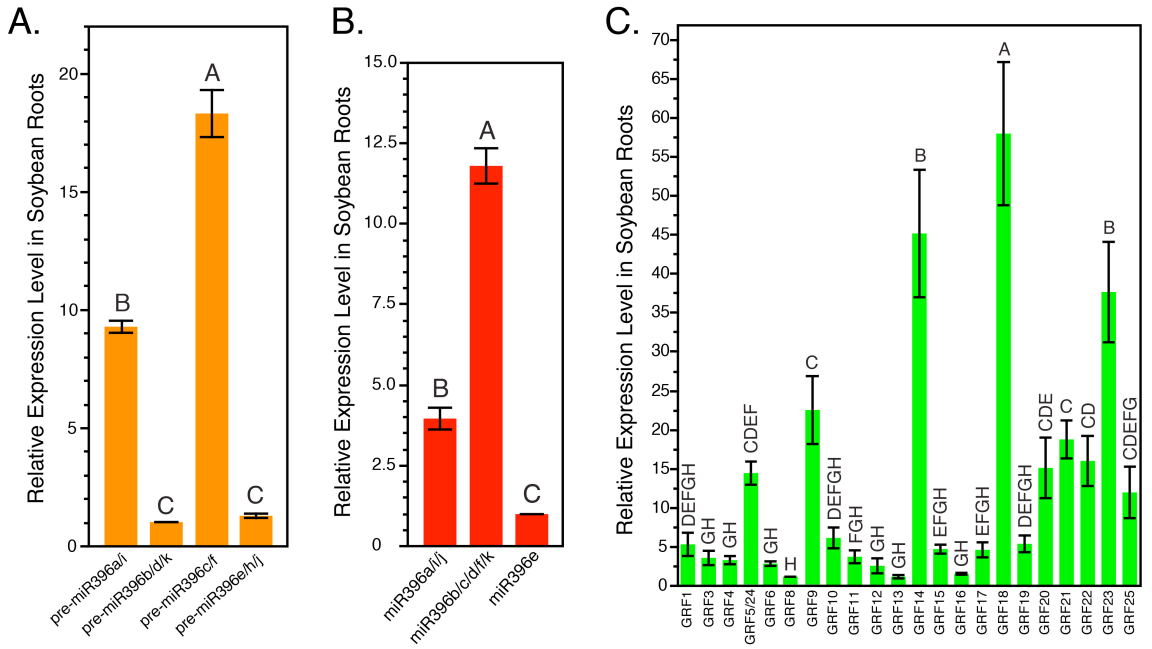



**Fig. 2. Native expression of pre-miR396, miR396 and *GRFs* in young soybean roots.** **(A)** qRT-PCR analysis of pre-miR396 subfamilies. Expression levels are relative to pre-miR396b/d/k. **(B)** qRT-PCR analysis of miR396 molecules. Expression levels are relative to miR396e. **(C)** qRT-PCR analysis of *GRF1-25*. Expression levels are relative to *GRF8*. (A-C) Error bars represent ± one standard deviation from the mean. Significance groups are shown ($P < 0.05$).

determine whether soybean *GRFs* are potentially targeted by miR396, we collected the mRNA sequences for all twenty-five *GRF* genes from SoyBase (Table S1, Appendix F) and submitted them to the psRNATarget server (Dai and Zhao 2011). Putative miR396 target sites were found in all but one *GRF* mRNA sequence (*GRF25*) (Fig. S1, Appendix F). *GRF25* did not encode a WRC DNA binding domain where the miR396 target site is located (Fig. S1, Appendix F). These data indicated that the mRNAs transcribed from *GRF1-24* have the potential for post-transcriptional regulation by miR396.

Because *GRF* genes are known to be active in early developing tissues and organs (Omidbakhshfard et al. 2015), we tested the expression of all twenty-five *GRFs* in 10-day-old soybean roots using qRT-PCR. *GRF2* and *7* mRNAs were undetectable using as

many as 40 cycles, while the remaining *GRF* mRNAs showed variable expression levels (Fig. 2C). Also, *GRF5/24*, *9*, *14*, *18*, *20-23* and *25* mRNAs resulted in much greater expression levels compared to the other thirteen *GRFs* (Fig. 2C). These findings indicated that twenty-three *GRFs* are active with variable levels of expression in young, developing soybean roots, consistent with the expression of the *MIR396* family members, and suggested that a miR396-*GRF* regulatory system operates in roots.

### 7.2.3 Eleven *GRF* genes are upregulated during the syncytium formation phase

We next examined whether *GRFs* change expression in response to *H. glycines* infection, specifically during the syncytium formation phase. However, we first had to determine which time point corresponded to the syncytium formation phase during *H. glycines* infection of soybean roots in our infection system. For this analysis, we inoculated soybean roots with *H. glycines* and at 2, 4, 8, 14 and 20-days-post-inoculation (dpi), evaluated which life stage the majority of *H. glycines* juveniles were in (Fig. 3A). We determined 2-4-dpi as the migration phase, 5-13-dpi as the syncytium formation phase, and 14-20-dpi as the syncytium maintenance phase in our infection assay (Fig. 3A).

A qRT-PCR screen was performed on RNA isolated from the 8-dpi roots for all twenty-five *GRFs*. Interestingly, eleven *GRF* mRNAs resulted in significantly increased expression compared to mock (between 3- and 8.5-fold increases, $P < 0.01$), while twelve *GRF* mRNAs were unchanged and *GRF2* and *7* remained undetected (Fig. S2, Appendix F). The eleven upregulated *GRFs* were *GRF6*, *8*, *9-13*, *15-17*, and *19* (Fig. S2, Appendix

F). These results suggested that a network of eleven different *GRFs* appear to be involved in *H. glycines* syncytium formation.

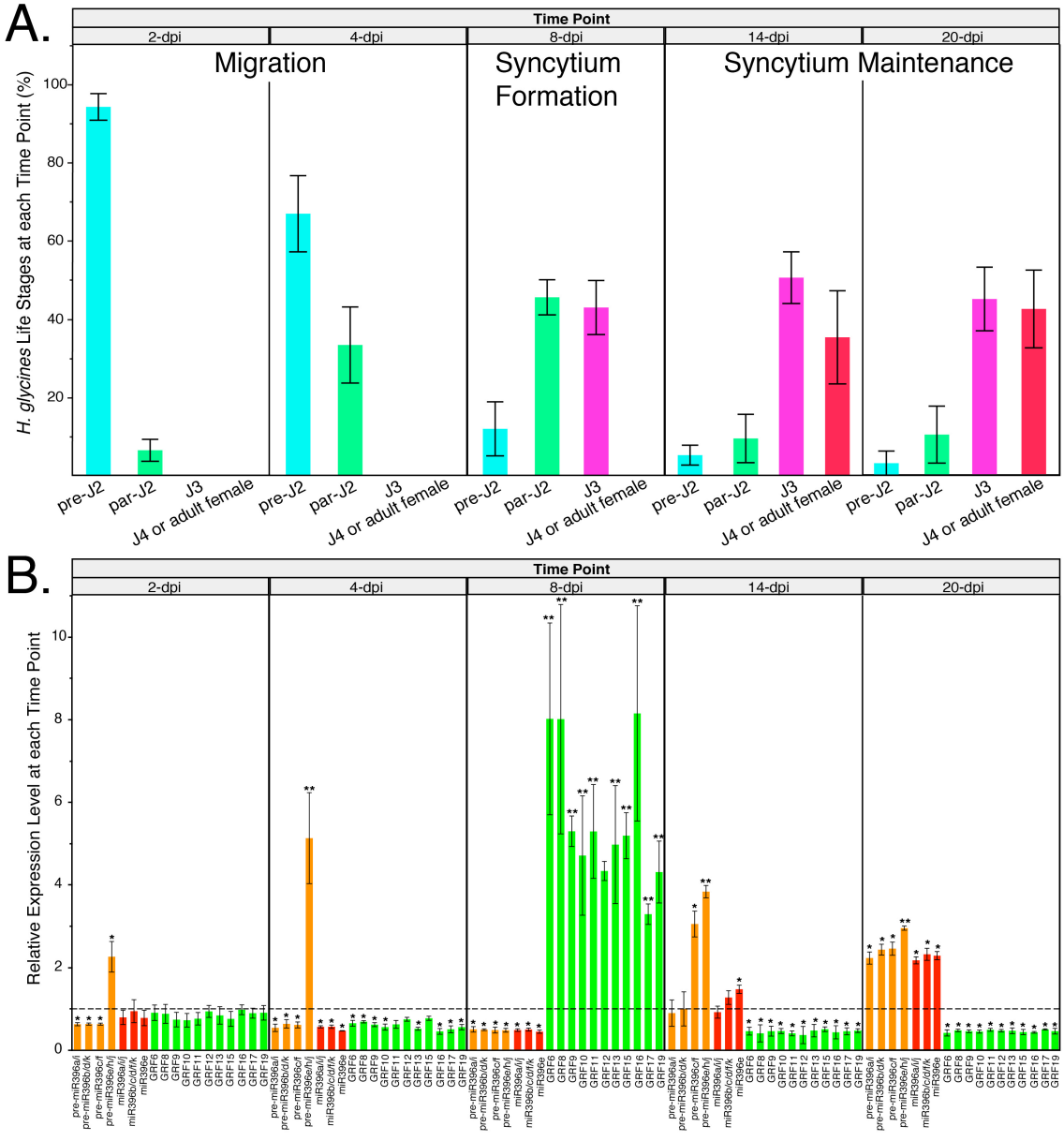



**Fig. 3. Time course expression analysis of the miR396-*GRF6/8-13/15-17/19* regulatory network during *H. glycines* infection. (A)** Assessment of syncytial phases in *H. glycines*-infected soybean roots ($n$ = 5 plants). At 2-dpi, an average of 94.1% of *H. glycines* were still in the pre-J2 stage while the few remaining were par-J2s. At 4-dpi, an average of 66.8% of *H. glycines* were in the pre-J2 stage and the remaining were par-J2s. Thus, we determined 2 and 4-dpi as early and late migration, respectively. By 8-dpi, an

average of 45.4% and 42.8% of *H. glycines* were in par-J2 or J3 stages, respectively, while the remaining few were pre-J2s. Thus, 8-dpi was designated as the syncytium formation phase. By 14- and 20-dpi, the majority of *H. glycines* were in late J3, J4 or adult female stages, and thus, these time points were designated as the syncytium maintenance phase. **(B)** Time course qRT-PCR analysis of pre-miR396 subfamilies, miR396 molecules, and the eleven *GRFs*. Expression levels are relative to mock; baseline expression is set to 1.0 and indicated with a dashed line. *, $P < 0.05$; **, $P < 0.01$. (A,B) Error bars represent ± one standard deviation from the mean.

### 7.2.4 The miR396-*GRF6/8-13/15-17/19* regulatory network delineates syncytium formation

Having determined that the *MIR396* gene family members are transcriptionally active in roots and that eleven *GRFs* are upregulated in response to *H. glycines,* it was of interest to examine the anticipated post-transcriptional silencing of these *GRFs* by miR396 during various stages of infection. We used qRT-PCR to quantify the expression of the four pre-miR396 subfamilies, the three miR396 molecules, and the eleven *GRFs* at 2, 4, 8, 14 and 20-dpi (Fig. 3B). With the exception of pre-miR396e/h/j, all pre-miR396 and miR396 as well as the *GRFs* showed no significant changes or only a slight downregulation during the migration time points (Fig. 3B). These slight downregulations are possibly a result of the substantial damage caused by the penetrating nematodes during migration. Interestingly, pre-miR396e/h/j showed significant upregulation during the migration time points, but this upregulation was not reflected by increased expression of the miR396 molecules, a possible indication of impaired miRNA maturation processing.

Strikingly, at 8-dpi all pre-miR396 and miR396 showed significant downregulation, and this downregulation was accompanied by significant upregulation of *GRFs* showing between 3- and 8.5-fold mRNA increases (Fig. 3B). Furthermore, at 14-

dpi all pre-miR396 and miR396 showed upregulation or no-significant changes, whereas all eleven *GRFs* were significantly downregulated over 2-fold (Fig. 3B). Moreover, at 20-dpi both pre-miR396 and miR396 were significantly upregulated, while all eleven *GRFs* remained significantly downregulated (Fig. 3B). The opposite expression patterns of miR396 and *GRFs* pointed to post-transcriptional silencing of the eleven *GRFs* by miR396 during *H. glycines* infection. These results indicated that the miR396-*GRF6/8-13/15-17/19* regulatory network delineates *H. glycines* syncytium formation.

### 7.2.5 *GRF6*, *8*, *9-13*, *15-17,* and *19* are post-transcriptionally regulated by miR396 in the *H. glycines* syncytium

To determine whether the gene expression changes of the *GRFs* are the results of their post-transcriptional regulation by miR396 during *H. glycines* infection, we performed a 5' RNA ligase-mediated (RLM)-RACE assay on the 14-dpi RNA (see Fig. 3B). Cloning and sequencing of the RLM-RACE clones indicated that the cleavage of all eleven *GRF* transcripts occurred within their miR396 target sites between positions 10 and 11 (Fig. 4, 10/10 clones). These results are consistent with previous reports for Arabidopsis *GRFs* (Hewezi et al. 2012; Jones-Rhoades and Bartel 2004) and confirmed that *GRF* mRNAs are post-transcriptionally regulated by miR396 in *H. glycines*-infected soybean roots.

Then, to confirm whether the eleven *GRFs* are post-transcriptionally regulated specifically in the syncytium, we examined the spatiotemporal expression patterns of these eleven *GRF* genes using promoter:GUS fusion assays. We generated transgenic hairy roots expressing promoter:GUS fusion constructs for the eleven *GRF* genes and all

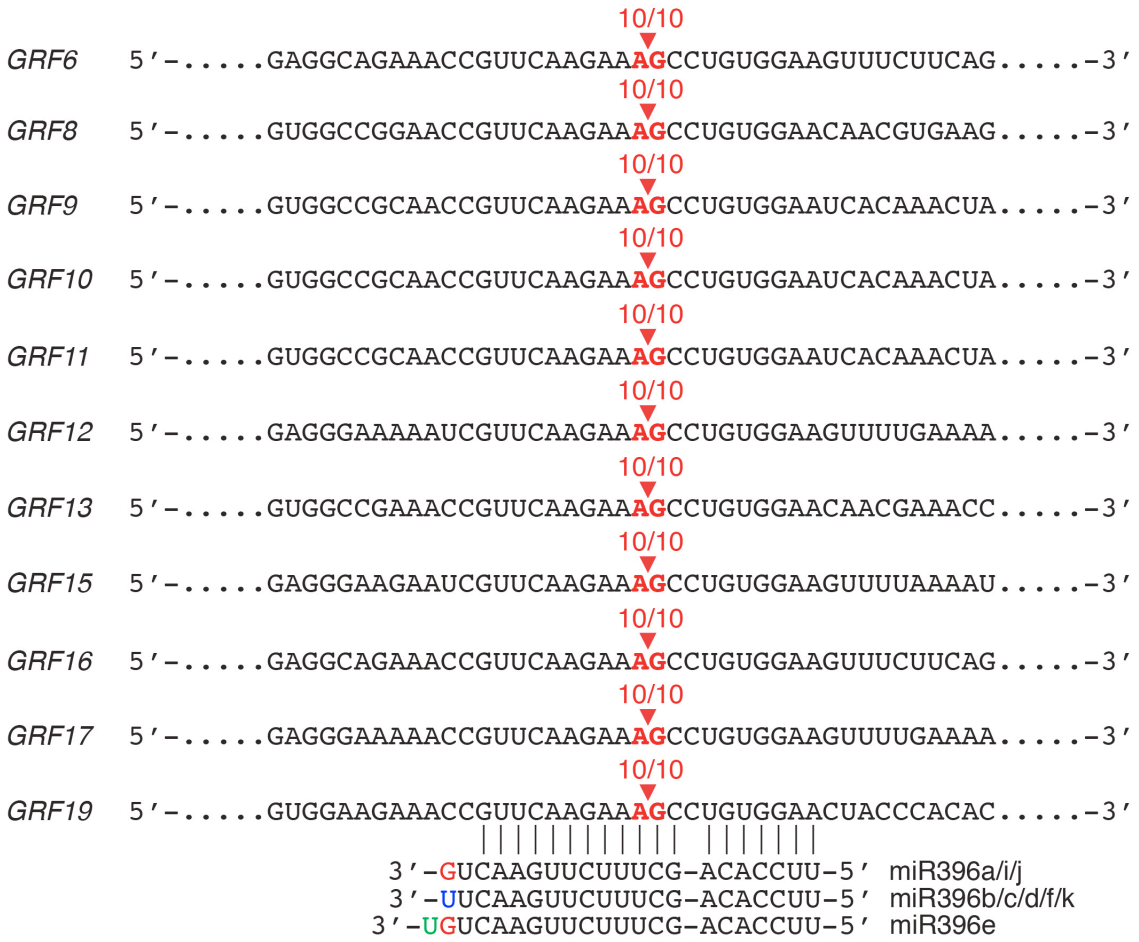
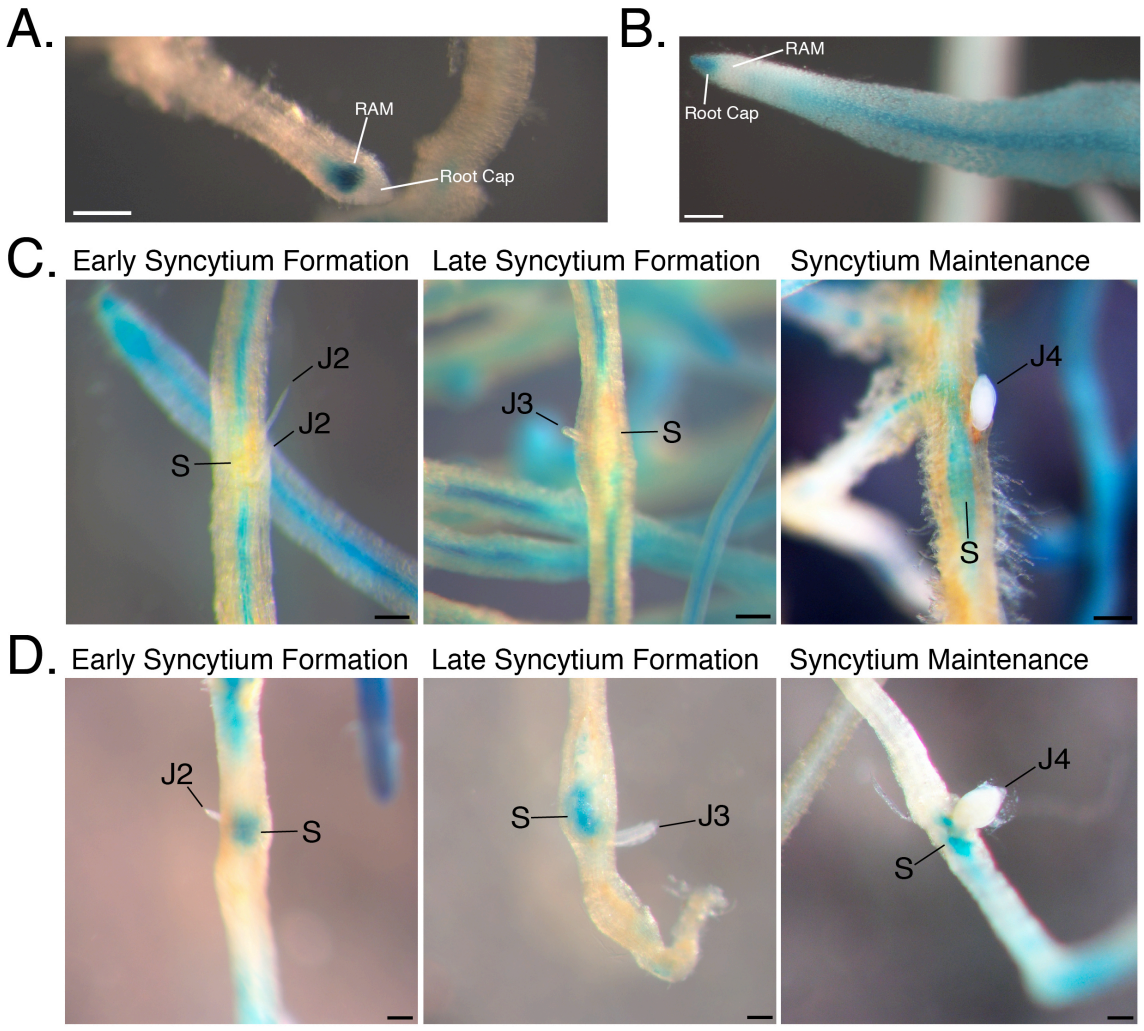



**Fig. 4. miRNA cleavage assays for *GRF6*, *8-13*, *15-17* and *19*.** Ten different clones for each *GRF* degradation product were analyzed by DNA sequencing. The number of clones that resulted in the indicated cleavage positions within the miR396 target sites is indicated.

nine functional *MIR396* genes. The histochemical localization of GUS activity directed by these promoters was assayed under both non-infected and *H. glycines*-infected conditions. Under non-infected conditions, the promoters of all *GRFs* produced strong GUS staining within the root apical meristem (RAM) and weak or undetectable activities in the vascular cylinder (Fig. 5A,E). Also, all *PMIR396:GUS* hairy roots showed consistent GUS activity in the vascular cylinder and in the root cap, but not in the RAM (Fig. 5B,E).




E.

| Summary of *MIR396* promoter activities | | | | | | |
|---|---|---|---|---|---|---|
| Promoter | RAM | Root Cap | VC near RAM | Syncytium | | |
| | | | | J2 | J3 | J4 |
| *MIR396a* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396b* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396c* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396d* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396e* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396f* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396i* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396j* | OFF | ON | ON | OFF | OFF | ON |
| *MIR396k* | OFF | ON | ON | OFF | OFF | ON |

| Summary of *GRF* promoter activities | | | | | | |
|---|---|---|---|---|---|---|
| Promoter | RAM | Root Cap | VC near RAM | Syncytium | | |
| | | | | J2 | J3 | J4 |
| *GRF6* | ON | OFF | OFF | ON | ON | ON |
| *GRF8* | ON | OFF | OFF | ON | ON | ON |
| *GRF9* | ON | OFF | OFF | ON | ON | ON |
| *GRF10* | ON | OFF | OFF | ON | ON | ON |
| *GRF11* | ON | OFF | OFF | ON | ON | ON |
| *GRF12* | ON | OFF | OFF | ON | ON | ON |
| *GRF13* | ON | OFF | OFF | ON | ON | ON |
| *GRF15* | ON | OFF | OFF | ON | ON | ON |
| *GRF16* | ON | OFF | OFF | ON | ON | ON |
| *GRF17* | ON | OFF | OFF | ON | ON | ON |
| *GRF19* | ON | OFF | OFF | ON | ON | ON |

**Fig. 5. GUSPlus histochemical analyses for *MIR396* and *GRF* promoters in non-infected and *H. glycines*-infected soybean roots. (A)** Native activity of *GRF* promoters. **(B)** Native activity of *MIR396* promoters. **(C)** Activity of *MIR396* promoters within the

syncytium. **(D)** Activity of *GRF* promoters within the syncytium. (A-D) Scale bars equal 0.5-mm. **(E)** Summary for all *MIR396* and *GRF* promoters.

Under *H. glycines*-infected conditions, all *MIR396* promoters showed clear downregulation in the syncytia induced by the second-stage (J2) and third-stage juvenile (J3) nematodes but became very active in the syncytia induced by fourth-stage (J4) nematodes (Fig. 5C,E). In contrast, all *GRF* promoters showed sustained upregulation in the syncytia induced by the J2, J3, and J4 nematodes (Fig. 5D,E). Noteworthy, no noticeable changes in *MIR396* and *GRF* promoter activities were observed in *H. glycines*-infected roots anywhere other than syncytia. These expression patterns were in strong support of the post-transcriptional regulation of *GRFs* by miR396 in the syncytium specifically during the later stage of infection.

### 7.2.6 Overexpression of pre-miR396 in soybean roots causes an *EXTRA HAIRY* phenotype and reduces *H. glycines* development to adult females

All unique, functional pre-miR396 were overexpressed in transgenic soybean hairy roots to determine whether interfering with the miR396-*GRF6/8-13/15-17/19* regulatory network would modify susceptibility to *H. glycines*. The constitutive soybean *polyubiqutin* promoter (Hernandez-Garcia et al. 2009) was used for overexpression (Fig. 6A). Transgenic soybean cv. Williams 82 roots with high pre-miR396 overexpression (Fig. 6B), as determined via qRT-PCR, were selected for phenotyping.

All pre-miR396 overexpression roots resulted in abnormal developmental phenotypes (Fig. 6C). Substantially more lateral roots were present within the first 2.5-cm from the root tips compared to vector control (Fig. 6D), and the phenotype was called

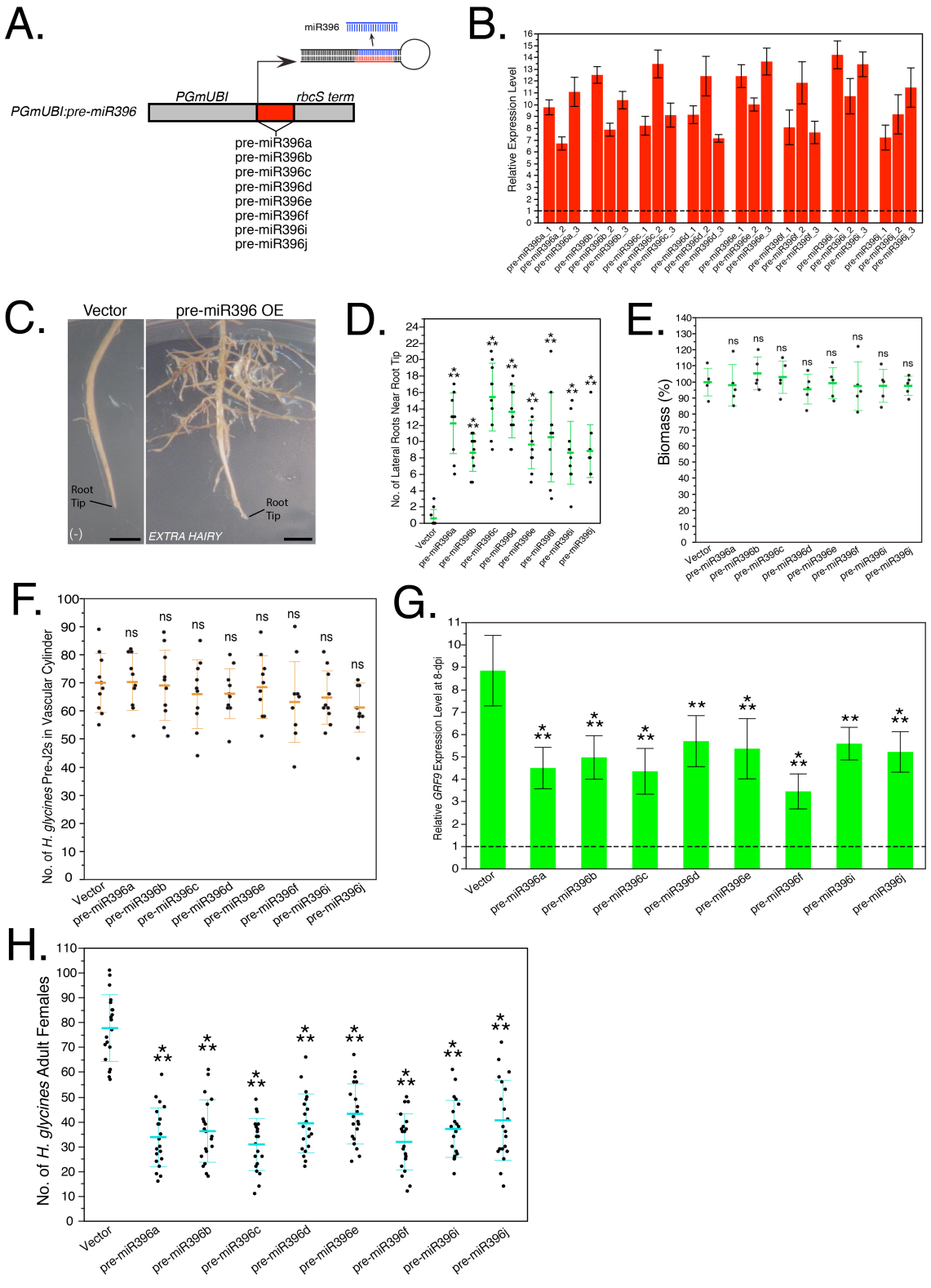



**Fig. 6. Overexpression of pre-miR396 in soybean roots. (A)** Overexpression constructs. **(B)** qRT-PCR analysis on transgenic pre-miR396-overexpressing roots, three

events per construct. Expression levels are relative to vector control; baseline expression is set to 1.0 and indicated with a dashed line. **(C)** *EXTRA HAIRY* developmental phenotype caused by pre-miR396 overexpression. Scale bars equal 0.5-cm. **(D)** Comparisons between the number of lateral roots within 2.5-cm from the root tips for pre-miR396-overexpressing and vector control roots ($n$ = 10). **(E)** Comparisons between biomasses of pre-miR396-overexpressing and vector control roots ($n$ = 5). Biomasses were measured as the percentage of dry root weight compared to vector control; vector control mean was set to 100%. (D,E) Data are representative of three independent experiments. **(F)** Comparisons between the number of *H. glycines* pre-J2s within the vascular cylinder of pre-miR396-overexpressing and vector control roots ($n$ = 10). (**G)** qRT-PCR analysis of *GRF9* in *H. glycines*-infected pre-miR396-overexpressing and vector control roots at 8-dpi. Expression levels are relative to mock for each construct; baseline expression is set to 1.0 and indicated with a dashed line. **(H)** Comparisons between the number of *H. glycines* adult females that developed on pre-miR396-overexpressing and vector control roots ($n$ = 20). (F,H) Data are representative of two independent experiments. (D-H) Error bars represent ± one standard deviation from the mean. **, $P < 0.01$; ***, $P < 0.001$; ns, not significant; all are statistically compared to vector control.

*EXTRA HAIRY*. On the other hand, the overall growths of pre-miR396 overexpression roots were not affected and root biomasses were the same as vector control (Fig. 6E), and there were no differences in the number of *H. glycines* penetrating J2s that infected them (Fig. 6F).

Also, all pre-miR396 overexpressing *H. glycines*-infected roots resulted in significantly weaker inductions of *GRFs* (i.e., *GRF9*) during syncytium formation compared to vector control (Fig. 6G). *GRF9* was selected as a representative for the other *GRFs* due to its high native expression in roots (Fig. 2C) and high upregulation during syncytium formation (Fig. 3C). Remarkably, all pre-miR396 overexpression roots resulted in highly significant reductions in the number of *H. glycines* adult females compared to vector control (Fig. 6H). Therefore, these results indicated that overexpression of all functional pre-miR396 reduces susceptibility to *H. glycines* by

inhibiting nematode development to the adult female stage, presumably by interfering with syncytium formation/function.

### 7.2.7 RNAi against a conserved region in *GRFs* phenocopies pre-miR396 overexpression

We next examined whether inactivation of *GRFs* would phenocopy the reduced nematode susceptibility observed from pre-miR396 overexpression. An RNA interference (RNAi) hairpin construct was generated that targeted a region in *GRF9* (nt 1-333 of the coding DNA sequence; CDS) that is conserved in the majority of *GRFs* (Fig. S3, Appendix F) and unique from other soybean sequences. *GRF9i*$_{1\text{-}333}$ was placed under the transcriptional control of the *GmUBI* promoter (Fig. 7A), and transgenic events that were determined to express *GRF9i*$_{1\text{-}333}$ via RT-PCR were selected for phenotyping.

All *GRF9i*$_{1\text{-}333}$ soybean roots phenocopied the *EXTRA HAIRY* phenotype that was observed for pre-miR396 overexpression roots (Fig. 7B) resulting in similar numbers of lateral roots within the first 2.5-cm from the root tip (Fig. 7C). Also, as observed for pre-miR396 overexpression roots, *GRF9i*$_{1\text{-}333}$ roots showed no difference in overall growth, root biomass (Fig. 7D) or in the number of *H. glycines* penetrating J2s that infected them compared to vector control (Fig. 7E).

Furthermore, the induction of *GRFs* (i.e., *GRF9*) during syncytium formation was significantly reduced in the *GRF9i*$_{1\text{-}333}$ roots (Fig. 7F) and resulted in a highly significant reduction in the number of *H. glycines* adult females (Fig. 7G). These results indicated that RNAi-mediated downregulation of *GRFs* results in both *EXTRA HAIRY*

developmental and reduced susceptibility phenotypes caused by pre-miR396 overexpression.

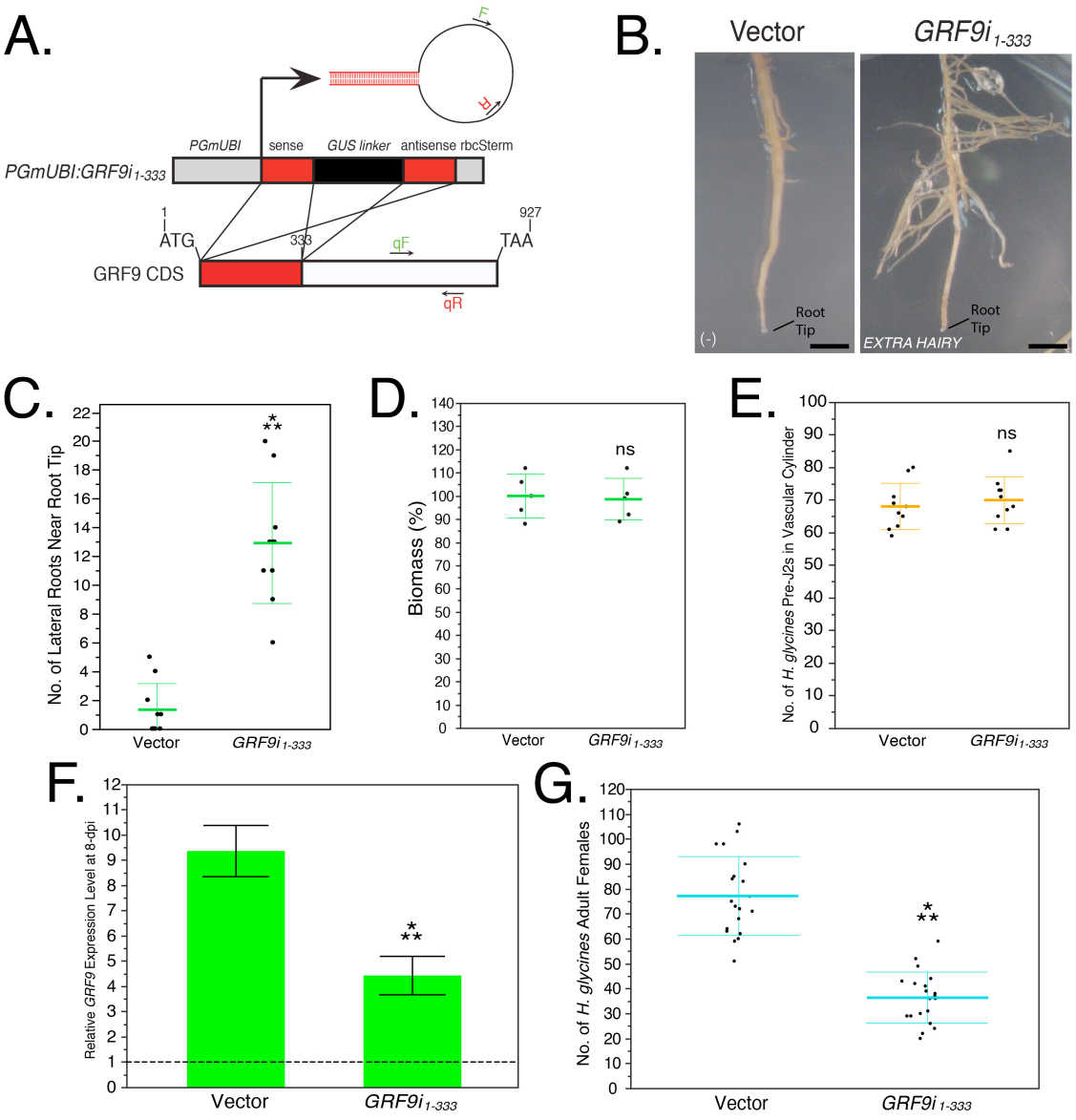



**Fig. 7. RNAi of nucleotides 1-333 of *GRF9* in soybean roots. (A)** RNAi construct. Annealing sites for the primers used for RT-PCR diagnosis of transgene expression (F and R), and qRT-PCR analysis of *GRF9* (qF and qR) are indicated. **(B,C)** *EXTRA HAIRY* developmental phenotype for $GRF9i_{1\text{-}333}$, as in (Fig. 6C,D). **(D)** Comparison between biomasses of $GRF9i_{1\text{-}333}$ and vector control roots, as in (Fig. 6E). **(E)** Comparisons between the number of *H. glycines* pre-J2s within the vascular cylinder of $GRF9i_{1\text{-}333}$ and vector control roots, as in (Fig. 6F). **(F)** Comparison between *GRF9* relative expression

levels in *H. glycines*-infected *GRF9i*$_{1\text{-}333}$ and vector control roots at 8-dpi, as in (Fig. 6G). **(G)** Comparisons between the number of *H. glycines* adult females that developed on *GRF9i*$_{1\text{-}333}$ and vector control roots, as in (Fig. 6H).

### 7.2.8 Overexpression of a miR396-resistant mutant of *GRF9* also phenocopies pre-miR396 overexpression

Finally, we tested whether overexpression of *GRFs* that are not under post-transcriptional control by miR396 would result in a perturbation of nematode susceptibility as was observed in the Arabidopsis-*H. schachtii* model (Hewezi et al. 2012). To test this hypothesis, a miR396-resistant, synonymous mutant of *GRF9* (*rGRF9*) with a modified miR396 target site was generated and placed under *PGmUBI* transcriptional control (Fig. 8A). Transgenic soybean roots determined to overexpress *rGRF9* at high levels via qRT-PCR were selected for phenotyping (Fig. 8B).

Once again, *rGRF9* overexpressing roots phenocopied the *EXTRA HAIRY* developmental phenotype observed for pre-miR396 overexpression and *GRF9i*$_{1\text{-}333}$ transgenic roots (Fig. 8C,D). Also, this manipulation did not alter the overall growth of these roots, the biomass (Fig. 8E) or the number of penetrating *H. glycines* J2s compared to vector control (Fig. 8F). However, overexpression of *rGRF9* in soybean roots resulted in significantly reduced numbers of *H. glycines* that developed to adult females compared to vector control (Fig. 8G). Thus, overexpression of *rGRF9*, like *GmGRF9i*$_{1\text{-}333}$, phenocopied the reduced susceptibility phenotype caused by pre-miR396 overexpression. Collectively, these *in vivo* studies indicated that homeostasis in the miR396-*GRF6/8-13/15-17/19* regulatory network is required for efficient *H. glycines* development to adult females.

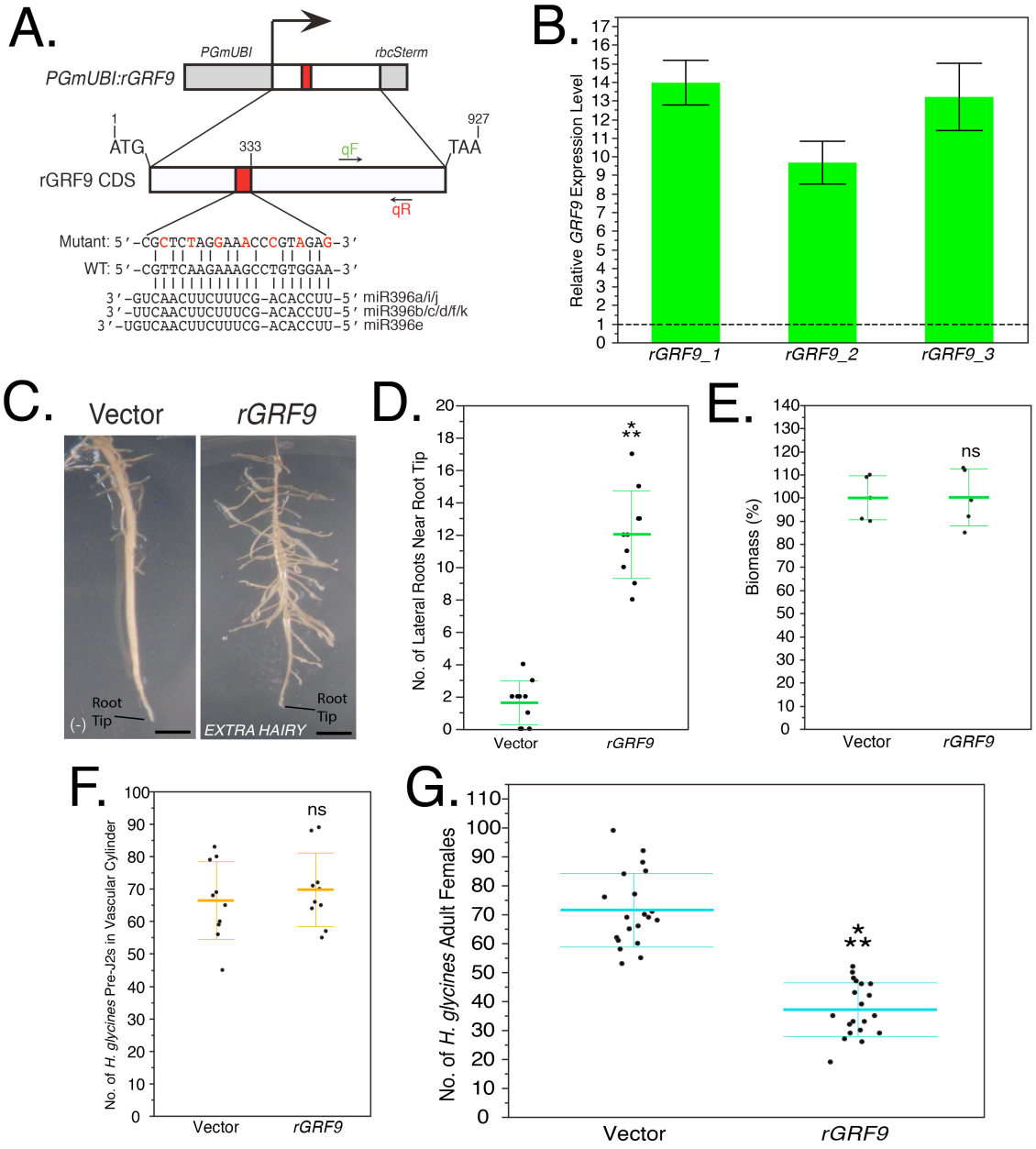



**Fig. 8**. **Overexpression of a miR396-resistant mutant of *GRF9* in soybean roots. (A)** Overexpression construct for miR396-resistant *GRF9* (*rGRF9*). An illustration of the synonymous mutations introduced into the miR396 target site is shown. Annealing sites for the primers used for qRT-PCR analysis of *GRF9* levels are shown. **(B)** qRT-PCR analysis of *GRF9* in transgenic *rGRF9*-overexpressing roots. Expression levels are relative to vector control; baseline expression is set to 1.0 and indicated with a dashed line. **(C,D)** *EXTRA HAIRY* developmental phenotype for *rGRF9* overexpression, as in (Fig. 6C,D). **(E)** Comparison between biomasses of *rGRF9*-overexpressing and vector control roots, as in (Fig. 6E). **(F)** Comparisons between the number of *H. glycines* pre-J2s within the vascular cylinder of *rGRF9*-overexpressing and vector control roots, as in (Fig.

6F). **(G)** Comparisons between the number of *H. glycines* adult females that developed on *rGRF9*-overexpressing and vector control roots, as in (Fig. 6G).

## 7.3 Discussion

Compared to Arabidopsis and the closely related legume *Medicago trucatula*, the soybean genome contains at least five times as many *MIR396* genes and almost three times as many *GRF* genes (Table S1, Appendix F) (Griffiths-Jones et al. 2008; Omidbakhshfard et al. 2015). Expansion of these gene families in soybean likely occurred through whole-genome duplication (WGD) and retention (Zhao et al. 2015a). Given that *MIR396* and *GRFs* are important for numerous biological processes in plants (Omidbakhshfard et al. 2015), their retention is probably not surprising. Expansion likely provided the adaptive benefits of organizational and regulatory diversity (Lespinet et al. 2002). However, it is interesting that twenty-four out of twenty-five *GRF* genes retained miR396-target sites (Fig. S1, Appendix F), underscoring the significance of their post-transcriptional regulation (Fig. 4). Also of particular interest, though there are at least ten *MIR396* genes, only three miR396 molecules appear to be produced. 21-nt miR396a/i/j and miR396b/c/d/f/k (Fig. 1B) are conserved among dicots and monocots (Kozomara and Griffiths-Jones 2014), and thus, have relatively ancient origins. However, 22-nt miR396e (Fig. 1B) seems to have appeared in soybean and may provide an additional, organizational/regulatory benefit, as 22-nt miRNAs can trigger the production of secondary siRNAs from target mRNAs (Chen et al. 2010). Furthermore, one *MIR396* gene in soybean appears to be nonfunctional, *MIR396h*, as the miR396/miR396* duplex does not form within the stem of the *in silico* pre-miR396h stem-loop (Fig. 1A). It may

also be possible that *MIR396h* experienced neofunctionalization and might produce a different miRNA.

Twenty-three *GRFs* showed variable expression levels in young soybean roots, while only two *GRFs* were undetected (Fig. 2C). We compared these data with RNA-seq data that are available for soybean root tips at the soybean functional genomics database (SFGD) (Yu et al. 2014). Even though these RNA-seq data were obtained from root tips as opposed to whole roots, and the plants were grown under different conditions than in our experiments, we found that our qRT-PCR data were well correlated with the SFGD RNA-seq data (Fig. S4; $R^2 = 0.48$, $P < 0.001$). Thus, this observation validates the accuracy of both our qRT-PCR data for *GRF* expression in whole soybean roots and the SFGD RNA-seq data for root tips. Also, *MIR396* gene family members exhibited variable expression levels in young soybean roots (Fig. 2A,B), which is consistent with previous findings (Li et al. 2012). However, Li et al. (2012) did not mention changes in miR396 abundance in response to *H. glycines*, but in their study soybean plants were grown in *H. glycines*-infested soil and they only evaluated a single time point long after infection was established. We found that all functional pre-miR396 and miR396 are downregulated much earlier during the syncytium formation phase, and upregulated during the maintenance phase (Fig. 3B). Thus, our finding that expression patterns differ drastically at different time points during *H. glycines* infection probably explain why Li et al. (2012) did not mention miR396.

Eleven soybean *GRFs* are upregulated during the syncytium formation phase, which in our infection system was from 5- to 13-dpi (Fig. 3A), consistent with previous descriptions (Ithal et al. 2007a), while the other *GRFs* do not change (Fig. S2, Appendix

F). Microarray analysis was previously performed on laser capture microdissected *H. glycines* syncytia at 2, 5 and 10-dpi (Ithal et al. 2007b). However, this study did not present any data on *GRFs*, which likely is due to a number of factors. For example, the only genes that were analyzed in the latter study were those that first changed in expression at 2-dpi, and then those genes were subsequently analyzed at 5- and 10-dpi. Thus, it is very likely that at 2-dpi *GRFs* are not yet upregulated since the peak of the syncytium formation phase occurs later on. Other microarray analyses were also performed on *H. glycines*-infected, whole soybean roots, but again no data was presented on *GRFs* (Alkharouf et al. 2006; Ithal et al. 2007a). Lack of data presented for *GRFs* in these studies could have also been due to insufficient representation on the GeneChip (Ithal et al. 2007a; Ithal et al. 2007b) or cDNA (Alkharouf et al. 2006) arrays, or possibly a combination of other factors. Also, many other microarray and RNA-seq studies have been performed on *H. glycines*-infected soybean roots, but the changes that are presented in those studies are representative of resistant reactions. Noteworthy, our qRT-PCR data for *GRF* expression changes during *H. glycines* infection (Fig. 3B) are consistent with the previously published microarray and qRT-PCR data for *H. schachtii*-infected Arabidopsis roots (Hewezi et al. 2012; Szakasits et al. 2009).

In the RAM, stem cell progeny undergo rapid cell division to ensure that there are enough cells for proper growth, and these rapidly dividing cells are called the transit-amplifying cells (TACs) (Rodriguez et al. 2015). In Arabidopsis roots, miR396 is abundant in the root cap and stem cell niche (SCN) formed by the quiescent center (QC) and adjacent stem cell initials, while *GRFs* are abundant in TACs. GRFs promote rapid cell cycling within TACs, and miR396-mediated downregulation of GRFs results in

delayed cell cycling (Rodriguez et al. 2015). We found that soybean *MIR396* promoters are all active within the root cap (most likely columella cells and the SCN) and the vascular cylinder leading up to the RAM (Fig. 5B,C), and that *GRF* promoters are predominantly active within the RAM, most likely TACs (Fig. 5A,C). Thus, the function of the miR396-*GRF* regulatory network in soybean root development appears to be very similar to other plant species (Bazin et al. 2013; Rodriguez et al. 2015). Also, *GRFs* are upregulated in the syncytium during the formation phase concomitant with the downregulation of miR396. Conversely, during the syncytium maintenance phase, *GRFs* are post-transcriptionally downregulated by the de-repressed miR396 expression (Fig. 3B). Thus, soybean *GRFs* are regulated in the *H. glycines* syncytium by miR396 in parallel to the RAM, suggesting that GRFs might function to maintain rapid cell cycling in the forming syncytium for proper organ development (Engler and Gheysen 2013).

When plants are under high pathogen stress, resources are devoted towards defense responses, while growth is stunted and development is delayed. This phenomenon is known as the growth-defense tradeoff (Huot et al. 2014). GRFs have been implicated in various abiotic and biotic stress conditions (Casadevall et al. 2013; Hewezi et al. 2012; Kim et al. 2012; Li et al. 2010b; Liu et al. 2008), and regulate the expression of a wide range of genes involved in both developmental processes and defense responses (Liu et al. 2014). GRFs are thus hypothesized to coordinate the interactions between defense signaling and growth and developmental pathways (Liu et al. 2014). In this context, GRFs could be thought to promote growth by maintaining rapid cell cycles while simultaneously suppressing defense responses. We found that silencing *GRFs* either by miR396 overexpression or RNAi greatly alters soybean root development and reduces

susceptibility to *H. glycines* (Figs. 6 and 7), consistent with *GRFs* promoting developmental processes and suppressing defenses. Interestingly, however, overexpression of *rGRF9* also alters root development and reduces susceptibility to *H. glycines* similar to *GRF* silencing (Fig. 8). Thus, although *GRFs* are required for proper soybean root development and successful *H. glycines* infection, their precise expression levels, fine-tuned by miR396, are also required.

Feedback regulation of miRNAs by their transcription factor targets has been demonstrated in several studies (Gutierrez et al. 2009; Hewezi and Baum 2012; Marin et al. 2010; Wang et al. 2009; Wu et al. 2009; Yant et al. 2010). Also, overexpression of *rGRF1* and *3* in Arabidopsis not only downregulates miR396, but also downregulates other *GRFs* as well as wild-type *GRF1* and *3*, respectively (Hewezi and Baum 2012). Although some of the coordination between miR396 and *GRFs* can be explained through PLETHORA (Rodriguez et al. 2015) and TCP4 (Rodriguez et al. 2010) transcription factors, it is clear that *MIR396* and *GRFs* are downstream targets that are negatively regulated by *GRFs* (Hewezi and Baum 2012). This complex feedback loop ensures a precise transcriptional equilibrium. Thus, pre-miR396 overexpression, RNAi against *GRFs*, and *rGRF9* overexpression all resulting in similarly altered root development and reduced susceptibility to *H. glycines* underscores the likely importance of this complex feedback loop to maintain such an equilibrium in soybean. Future studies that analyze the expression changes in miR396 and *GRFs* in *rGRF9*-overexpression roots will provide a more complete picture of the necessary feedback regulations within the miR396-*GRF6/8-13/15-17/19* regulatory network during *H. glycines* infection.

In summary, we have conducted a translational research study to investigate whether a miR396-*GRF* regulatory system operates in the agronomically important interaction between *H. glycines* and soybean. Our results demonstrate that the miR396-*GRF6/8-13/15-17/19* regulatory network delineates syncytium formation and that interfering in the homeostasis of this network prevents efficient *H. glycines* development to adult females. As *H. glycines* is the most economically devastating soybean pathogen causing over US$1 billion in yield losses each year (Koenning and Wrather 2010), control strategies more effective than the conventional measures (Conley et al. 2011) are urgently needed. Thus, this network serves as a promising target to develop soybean plants with novel, synthetic resistance to *H. glycines*.

## 7.4 Experimental Procedures

### 7.4.1 Inoculation of whole plants

Soybean cv. Williams 82 seeds were surface sterilized with 10% sodium hypochlorite for 10-min and planted on seed germination paper (Anchor Paper). Ragdolls were incubated at 26˚C in the dark for 3-days. Seedlings were placed on Circular Steel Blue Seed Germination Blotter Paper (Anchor paper) dampened with MES buffered ddH20, pH 6.5 in a circle with the radical tips facing towards the center (10 seedlings per plate). Each radical was inoculated with 500 surface-sterilized *H. glycines* pre-J2s (Baum et al. 2000). Inoculated radicals were covered with dampened Blotter Paper, and infection chambers were incubated at 26˚C in the dark for 24-hr. Four inoculated seedlings were acid fuchsin stained for *H. glycines* (Hussey 1985) to ensure good infection. Infected

seedlings were rinsed and placed back into ragdolls and incubated in a Percival growth chamber at 26˚C with a 14:10-hr light-dark cycle.

### 7.4.2 *in silico* analyses

Pre-miR396 sequences were blastn-searched against the soybean genome at SoyBase using default parameters. Pre-miR396 stem-loops were modeled *in silico* using the Mfold Web Server (Zuker 2003) with default settings. All GRF CDS sequences were submitted to the psRNATarget server (Dai and Zhao 2011) with default parameters along with all miR396 molecules to evaluate for putative miR396 target sites.

### 7.4.3 Phylogenetic analysis

Multiple sequence alignments (MSAs) were generated with Clustal using default parameters in MEGA6 (Tamura et al. 2011). Poorly aligned regions were removed. Model selection analysis was performed in MEGA6 using default parameters to obtain the best-scoring model of nucleotide substitution. Phylogenetic analysis was performed in MEGA6 using bootstrapped ML estimation with 100 bootstrap replications. Reported is the best scoring ML phylogenetic tree with bootstrap values indicated on the corresponding nodes.

### 7.4.4 Assessment of syncytial phases

Roots from five plants per time point were acid fuchsin stained for *H. glycines* (Hussey 1985). For each plant, 100 *H. glycines* were observed with a Stereo Microscope (Zeiss) and each life stage was recorded. The average percentage of each life stage was

calculated for each time point. This method was also used to compare the number of *H. glycines* penetrating J2s that infected (i.e., reached the vascular cylinder) for experimental and vector control roots.

### 7.4.5 RNA isolation and cDNA synthesis

Total RNA was isolated from 50-mg of ground root tissue using the NucleoSpin Kit (Clontech). Yields and purity were assessed with a NanoDrop, and integrity with agarose gel electrophoresis. Total RNA was polyadenylated and reverse transcribed using the Mir-X miRNA Kit (Clontech).

### 7.4.6 qRT-PCR

qRT-PCR was performed with iQ SYBR Green (Bio-Rad) on an iCycler iQ Real-Time PCR Detection System (Bio-Rad). Protocol: 95˚C for 3-min, 40 cycles of 95˚C for 15-sec and 60˚C for 30-sec. Universal mRQ reverse primer (Mir-X miRNA Kit) was used along with miR396-specific oligonucleotides as forward primers. Forward primers specific to each pre-miR396 subfamily were used with mRQ reverse. Pre-miR396 had to be quantified as subfamilies because primers that attempted to quantify each subfamily member individually resulted in much lower primer efficiencies. miR396-specific forward primers included two adenine nucleotides on the 3'-ends to ensure binding to the poly(T) region of miR396 cDNAs and not to pre-miR396 (Gutierrez et al. 2009). U6 (Mir-X miRNA Kit) was used as a calibrator for normalization. For *GRFs*, RNA levels were normalized to *GmUBQ3* (GenBank: D28123.1). Amplification specificities were confirmed for all by melting curve analysis and agarose gel electrophoresis. Melting

curve analysis protocol: 95˚C for 1-min, 55˚C for 10-sec and a slow temperature ramp from 55-95˚C. ddH20 and total RNA samples were included as negative controls, and no amplification was obtained. Three biological replicates and four technical replicates were always used. Relative changes in gene expression levels were quantified using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001). Single statistical comparisons were made using the t-test, and multiple comparisons by ANOVA and Tukey-Kramer HSD post-hoc test in JMP Pro 11.

### 7.4.7 miRNA cleavage assays

miR396 cleavage sites were mapped with the FirstChoice RLM-RACE Kit (Ambion). 14-dpi total RNA was poly(a)-selected with Dynabeads (Thermo) and ligated to the 5'-RACE RNA adaptor without calf intestine alkaline phosphatase treatment. cDNA synthesis was performed using *GRF*-specific outer primers. Subsequent steps followed the manufacturer's instructions. RLM-RACE products were cloned into pGEM-T Easy (Promega) and sequenced at Iowa State University.

### 7.4.8 Vector construction

For promoter constructs, soybean Williams 82 genomic (g)DNA was isolated from a leaf of a 3-week-old plant according to (Blin and Stafford 1976). 1.4-2.3-Kbp of upstream regulatory DNA sequence in SoyBase was cloned for each promoter construct. PCR amplification was performed with Platinum *Taq* (Invitrogen). PCR products were cloned into pGEM-T Easy and restriction-digest-cloned into p4305.1 (GenBank: KT954098) (restriction enzyme sites are included on the primer sequences in Table S2,

Appendix F). For pre-miR396 overexpression constructs, pre-miR396 were PCR-amplified from soybean gDNA using Platinum *Taq* with primers exactly 20-nt 5' and 3' to the pre-miR396 sequences in SoyBase. PCR products were cloned into pGEM-T Easy and restriction-digest-cloned into the pG2XPRESS derivative of pG2RNAi2 (GenBank: KT954097); the *GUS linker* was restriction-digested out. For RNAi, we PCR-amplified nt 1-333 of the GRF9 CDS (*GRF9i*$_{1\text{-}333}$) from soybean cDNA with Platinum *Taq*. PCR products were cloned into pGEM-T Easy and restriction-digest-cloned into pG2RNAi2 sense and antisense sites. For *rGRF9* overexpression, we PCR-amplified the *GRF9* CDS from cDNA, and synonymous mutations were introduced in the miR396 target site by overlap extension PCR (Ho et al. 1989). The *rGRF9* PCR product was cloned into pGEM-T Easy and restriction-digest-cloned into pG2XPRESS. All vectors were sequenced, and then transformed into *Agrobacterium rhizogenes* strain K599.

### 7.4.9 Hairy root nematode infection assays

Transgenic hairy roots were generated and inoculated with surface-sterilized *H. glycines* (500 pre-J2s per root tip) according to (Baum et al. 2000; Liu et al. 2012b). Each replicate consisted of a maintenance plate of 10 infected parent roots. Statistical comparisons were made with the t-test in JMP Pro 11.

### 7.4.10 GUSPlus histochemical staining

Transgenic hairy roots were inoculated with surface-sterilized *H. glycines* in 6-well plates (350 pre-J2s per well) similar to (Baum et al. 2000). Infected and noninfected roots were removed from the solid media, the solid media was removed, and then roots

were placed back into the empty 6-well plates and subjected to histochemical staining for GUSPlus according to (Vitha et al. 1995). Stained roots were mounted in ddH20 or glycerol and observed with a Stereo Microscope. Images were taken with an AxioCam HR 13 Megapixel Camera (Zeiss).

## 7.5 Authors' contributions

JBN conceived and designed research, conducted experiments, and analyzed data. TH and TJB co-wrote the manuscript with JBN. TJB supervised the work.

# CHAPTER 8. GENERAL CONCLUSIONS

Nematode adaptation to plant parasitism within the suborder Hoplolaimina involved several major genomic and cellular changes. A principle genomic change was the horizontal acquisition of many genes from bacteria (e.g., genes that encode cell wall modifying proteins and enzymes for vitamin B biosynthesis and salvage pathways). We determined that several additional genes [i.e., *GLAND1 GCN5-related N-acetyltransferases* (*GNATs*), *invertases* and *chorismate mutases*] were also acquired from bacteria in these nematodes, and presently encode candidate effectors. We further showed that the bacterial donors of these genes inhabited the same niche as the Hoplolaimina nematodes (i.e., the rhizosphere) further strengthening the support for the HGT hypothesis. Additionally, we found that *GNATs* homologous to *GLAND1s*, and *invertases*, were the subjects of numerous HGTs from diverse bacteria to many different eukaryotes and archaea. These latter findings demonstrated that these genes likely provided important evolutionary advantages not just for these nematodes, but many different taxa.

Despite the prevalence of HGT in Hoplolaimina nematodes, the mechanism(s) for HGT within this lineage was completely unknown. However, in other eukaryotes, transposable elements (TEs) had been hypothesized to contribute to HGT based on their enrichment near HGT genes. We demonstrated that in the genomes of at least some Hoplolaimina nematodes (i.e., cyst nematodes *H. glycines* and *G. pallida*, and root-knot nematode *M. incognita*) HGT genes are significantly associated not with TEs per say, but with *Cer* retroviruses. Previous findings demonstrated that *Cer* retroviruses had the

potential to be active within the germ line of, and to be infectious to, Hoplolaimina nematodes. In combination with the potential for sequence homology-independent, intermolecular strand-transfers, these potential features of *Cer* retroviruses might have allowed for the capture of foreign DNA, movement into the germ line, and integration into the recipient genome. We therefore proposed the tempting hypothesis that *Cer* retroviruses might have contributed to HGT within this lineage.

A major cellular change in Hoplolaimina nematodes was the evolution of the elaborate, unicellular, secretory glands associated with the esophagus. These esophageal glands evolved into the major apparatuses for effector production and secretion. In previous projects, many candidate or *bona fide* effectors were determined to be produced specifically in the esophageal glands. These projects mostly focused on the agronomically important species *H. glycines*, *G. pallida*, *G. rostochiensis* and *M. incognita*. The protein sequences of most of these candidate effectors were pioneers (i.e., no homologs could be found in sequence databases) and all contained N-terminal signal peptides for classical secretion. In our work, we identified eighteen new candidate effectors in *H. glycines* produced specifically in the esophageal glands that were missed in previous projects. The majority of these protein sequences were once again pioneers, further demonstrating the predominance of pioneer effectors in these nematodes. Also, all eighteen of these new candidate effectors contained signal peptides for secretion. Moreover, of the more than one hundred candidates without signal peptides that were included in our tests, none were found to be expressed in the esophageal glands. These findings underscore the classical secretory pathway as the predominant mode of secretion from the esophageal glands. Noteworthy, the diverse expression profiles of these new

candidate effectors in the esophageal glands throughout parasitic stages suggest their timely importance for parasitism, and thus serve as potential targets for control measures.

Nematodes cause extensive damage and modification to host plant root tissues and cells during infection. Intracellular migration and formation of syncytia qualify the cyst nematodes as being among the most destructive Hoplolaimina. It is therefore expected that cyst nematodes largely suppress the immune systems of their plant hosts in order to sustain such prolonged interactions. However, very few *H. glycines* effectors have been characterized directly, revealing a major gap in the understanding of how this nematode remains such a successful soybean pathogen. We showed that a *H. glycines* effector called HgGLAND18 identified by us is strongly expressed in the dorsal gland during each stage of the nematode's life cycle. HgGLAND18 suppresses both canonical pattern- and effector-triggered immunity pathways, and its immunosuppression is correlated with efficiency of localization to plant cell nucleoli. Moreover, we determined that HgGLAND18 is essential for *H. glycines* virulence. Thus, these findings documented the first connection between the nucleolus and plant immunity to nematodes, and presented HgGLAND18 as an attractive target for control measures against *H. glycines*.

An intriguing feature of HgGLAND18 is the N-terminal sequence immediately after the signal peptide that we called the CSP-like domain. The CSP-like domain contains marginal, but significant sequence similarity to domains RI, RR and RII+ from CSPs of three Asian primate malarias (i.e., *Plasmodium fieldi*, *P. simiovale* and a *P. vivax*-like species). These similarities cannot be found in any other nematode sequence, including the GLAND18 ortholog from *H. schachtii*, the sister species to *H. glycines*, and the HgGLAND18 paralogous effector HgGLAND8. Also, extensive searches of sequence

databases determined that sequences from other organisms are very unlikely to contain such similarities. Interestingly, *Plasmodium* CSPs, like HgGLAND18, are immunosuppressive effectors in their vertebrate hosts, and their functions largely depend on domains RI, RR and RII+. Remarkably, substitution of the respective domains from *P. fieldi* CSP for the CSP-like domain in HgGLAND18 resulted in full complementation of immunosuppression. We therefore concluded that the similarity between the CSP-like domain in HgGLAND18 and the CSP domains from the respective *Plasmodium* species represents a remarkable example of convergent evolution between two very distantly related pathogens. Moreover, convergence of these effector sequences is best explained by similar immunosuppressive functions in their respective host cells.

The sedentary endoparasites within Hoplolaimina (e.g., cyst and root-knot nematodes) form elaborate feedings sites within the vascular cylinder of infected roots. Feeding site formation likely manifests in part through the coordinated functions of nematode effectors and their downstream host targets. A particular downstream host regulatory system that involves plant microRNA396 (miR396) and its *Growth-Regulating Factor* (*GRF*) targets was shown in the model cyst nematode interaction between *H. schachtii* and Arabidopsis to be necessary for syncytium formation. We performed a translational research project to investigate whether a similar regulatory system is necessary for syncytium formation in the agronomically important interaction between *H. glycines* and soybean. Reassuringly, this much-expanded regulatory network in soybean, which consists of nine functional *MIR396* and eleven *GRF* genes (i.e., the miR396-*GRF6/8-13/15-17/19* regulatory network), was largely conserved. For example, as found in the model pathosystem, all *MIR396* are downregulated and the targeted *GRFs*

are upregulated during the syncytium formation phase. During the switch to the syncytium maintenance phase, the opposite is observed; all *MIR396* are upregulated and *GRFs* are downregulated. Also, the downregulation of *GRFs* during syncytium maintenance occurs specifically from post-transcriptional regulation by miR396. Finally, interference in this regulatory network either through overexpression or RNAi knockdown experiments greatly reduces *H. glycines* development to the adult female stage, presumably as a consequence of altered syncytium formation/function. This regulatory network therefore represents an ideal target for engineering synthetic resistance against this important soybean pathogen.

The research included in this dissertation revealed numerous novelties for our understanding of the molecular biology and evolution of plant parasitism by nematodes. This basic knowledge will undoubtedly provide foundations for future studies geared towards the development of novel control measures against plant-parasitic nematodes. Additionally, several important findings were made for the agronomically important *H. glycines*-soybean pathosystem that could potentially be used directly for biotechnological application. These important findings include greatly reduced susceptibility through host-induced RNAi against important *H. glycines* genes such as *HgGLAND18*, and interference in the soybean miR396-*GRF6/8-13/15-17/19* regulatory network during syncytium formation, as detailed below.

There is currently much debate for whether transgenic expression of hairpin constructs in the host plant for RNAi against nematode genes works in the *H. glycines*-soybean pathosystem. We showed that a hairpin construct that utilizes a *GUSPlus* linker sequence in between two fragments of *HgGLAND18* in reverse complement orientation

reduces target transcripts in parasitic *H. glycines* with no observed off-target effects. We consistently observed between 50-75% reduction in *HgGLAND18* target transcripts. These specific reductions in *HgGLAND18* transcripts consistently reduced susceptibility by 40-70% in multiple, compatible *H. glycines*-soybean pathosystems (i.e., avirulent *H. glycines* on susceptible soybean and virulent *H. glycines* on resistant soybean). Therefore, our findings provide direct evidence that host-induced RNAi against important *H. glycines* genes such as *HgGLAND18* is a promising strategy for biotechnological control of *H. glycines* in soybean.

Interference in the soybean miR396-*GRF6/8-13/15-17/19* regulatory network, similar to host-induced RNAi against *HgGLAND18*, consistently resulted in around 40-70% reduced susceptibility to *H. glycines*. Three independent methods of interference resulted in similar outcomes. These methods included overexpression of each individual pre-miR396, RNAi against *GRFs*, as well as overexpression of a *GRF9* mutant with a synonymously mutated miR396 target site. The former two methods decreased the level of upregulation of *GRFs* during the syncytium formation phase, while the latter method eliminated miR396 control over *GRF9*. These findings indicated that precise miR396 control over *GRFs* is necessary for *H. glycines* development, presumably through an essential role in syncytium formation/function. Furthermore, though we consistently observed 30-60% of *H. glycines* that still developed to adult females, interference in this regulatory network can probably be optimized to increase the potency and specificity (see Appendix G for preliminary data on stable transgenic soybeans). Future optimization studies may ultimately determine the efficacy of this novel control measure for soybean's

most devastating pathogen. Furthermore, it will be interesting to see if this system can also be targeted to control for other agronomically important cyst nematodes.

# APPENDIX A. CHAPTER 2 SUPPLEMENTARY INFORMATION

**Table S1. List of candidate effectors identified by Gao et al. (2003) that were rediscovered from our gland cell library, and the number and corresponding IDs of our matching ESTs**

| ID | Accession | Matching ESTs from our gland cell library | No. of ESTs that aligned |
|---|---|---|---|
| 2A05 | AY028639 | hgg1c.pk011.j20.f hgg1c.pk011.j20 | 2 |
| 2B10 | AF273728 | hgg1c.pk014.i12 hgg1c.pk001.b13.f hgg1c.pk002.d9.f hgg1c.pk009.a19.f hgg1c.pk003.m4.f hgg1c.pk007.h21.f hgg1c.pk005.i8.f hgg1c.pk008.b20.f hgg1c.pk004.d12.f hgg1c.pk014.i12.f hgg1c.pk004.g4.f hgg1c.pk050.f11.f hgg1c.pk002.d9 hgg1c.pk008.p21 hgg1c.pk001.b13 hgg1c.pk007.h21 hgg1c.pk008.b20 hgg1c.pk005.i8 hgg1c.pk004.d12 hgg1c.pk014.d24 hgg1c.pk004.g4 hgg1c.pk003.m4 hgg1c.pk050.f11 hgg1c.pk009.a19 | 24 |
| 2D01 | AF469057 | hgg1c.pk014.l13 hgg1c.pk014.l13.f hgg1c.pk008.n5.f | 3 |
| 3D11 | AF468679 | hgg1c.pk006.i12 hgg1c.pk005.e11 hgg1c.pk005.e11.f hgg1c.pk006.i12.f hgg1c.pk008.b14.f hgg1c.pk015.k10 hgg1c.pk008.b14 | 7 |
| 3H07 | AF473831 | hgg1c.pk002.o20.f hgg1c.pk003.c8.f hgg1c.pk014.g8.f hgg1c.pk002.o20 hgg1c.pk049.a15 hgg1c.pk048.a15.f hgg1c.pk003.c8 hgg1c.pk048.a15 hgg1c.pk049.a15.f hgg1c.pk014.g8 | 10 |
| 4D06 | AF469063 | hgg1c.pk012.e9.f hgg1c.pk006.g12 hgg1c.pk006.g12.f hgg1c.pk014.k4 hgg1c.pk001.l20.f hgg1c.pk014.k4.f hgg1c.pk014.o19.f hgg1c.pk006.k18 hgg1c.pk001.d23 hgg1c.pk006.i16 hgg1c.pk007.m21 hgg1c.pk001.l20 hgg1c.pk005.c8 hgg1c.pk007.k10 hgg1c.pk002.b9 hgg1c.pk052.e13 hgg1c.pk052.e13.f hgg1c.pk006.k18.f hgg1c.pk007.m21.f hgg1c.pk005.c8.f hgg1c.pk007.k10.f hgg1c.pk006.i16.f hgg1c.pk002.b9.f hgg1c.pk001.d23.f | 24 |
| 4D09 | AF469061 | hgg1c.pk015.d6 hgg1c.pk015.d6.f hgg1c.pk015.c5.f | 3 |
| 4E02 | AF473826 | hgg1c.pk006.h9.f hgg1c.pk006.h9 | 2 |
| 4F01 | AF469059 | hgg1c.pk001.h10 hgg1c.pk001.h10.f hgg1c.pk052.k17.f hgg1c.pk052.k17 hgg1c.pk050.f12 hgg1c.pk050.f12.f hgg1c.pk008.f1 hgg1c.pk008.f1.f | 8 |
| 4G06 | AF469060 | hgg1c.pk002.o20.f hgg1c.pk003.c8.f hgg1c.pk014.g8.f hgg1c.pk002.o20 hgg1c.pk049.a15 hgg1c.pk048.a15.f hgg1c.pk003.c8 hgg1c.pk048.a15 hgg1c.pk049.a15.f hgg1c.pk014.g8 | 10 |
| 4G12 | AF473827 | hgg1c.pk014.i12 hgg1c.pk001.b13.f hgg1c.pk002.d9.f hgg1c.pk009.a19.f hgg1c.pk003.m4.f hgg1c.pk007.h21.f hgg1c.pk005.i8.f hgg1c.pk008.b20.f hgg1c.pk004.d12.f hgg1c.pk014.i12.f hgg1c.pk004.g4.f hgg1c.pk050.f11.f hgg1c.pk002.d9 hgg1c.pk008.p21 hgg1c.pk001.b13 hgg1c.pk007.h21 hgg1c.pk008.b20 hgg1c.pk005.i8 hgg1c.pk004.d12 hgg1c.pk014.d24 hgg1c.pk004.g4 hgg1c.pk003.m4 hgg1c.pk050.f11 hgg1c.pk009.a19 | 24 |
| 5D08 | AF473828 | hgg1c.pk008.l7.f hgg1c.pk008.l7 | 2 |
| 6E07 | AF473829 | hgg1c.pk050.m6.f hgg1c.pk011.g4 hgg1c.pk050.m6 | 3 |

Table S1 continued

| | | | |
|---|---|---|---|
| 6F06 | AY043224 | hgg1c.pk009.c16.f hgg1c.pk009.c16 | 2 |
| 8H07 | AF500024 | hgg1c.pk010.h19.f hgg1c.pk015.d15 hgg1c.pk015.d15.f hgg1c.pk010.f13.f | 4 |
| 10A07 | AF500021 | hgg1c.pk012.g22.f hgg1c.pk008.b16.f hgg1c.pk012.g22 hgg1c.pk008.b16 | 4 |
| 10C02 | AF500017 | hgg1c.pk006.h9.f hgg1c.pk006.h9 | 2 |
| 11A06 | AF500015 | hgg1c.pk002.j20 hgg1c.pk002.j20.f hgg1c.pk003.g23.f hgg1c.pk006.p14 hgg1c.pk006.p14.f | 5 |
| 12H04 | AF490244 | hgg1c.pk006.j7.f hgg1c.pk006.j7 | 2 |
| 13A06 | AF500020 | hgg1c.pk050.m6.f hgg1c.pk011.g4 hgg1c.pk050.m6 | 3 |
| 17C07 | AF520566 | hgg1c.pk009.c2.f hgg1c.pk009.c2 | 2 |
| 17G06 | AF490247 | hgg1c.pk005.l7.f | 1 |
| 18H08 | AF490248 | hgg1c.pk013.i3.f hgg1c.pk014.g18.f hgg1c.pk013.i3 hgg1c.pk014.g18 | 4 |
| 19C07 | AF490250 | hgg1c.pk012.m2 hgg1c.pk049.c24 hgg1c.pk003.p17 hgg1c.pk003.p17.f hgg1c.pk049.a24.f hgg1c.pk049.c24.f hgg1c.pk048.c24.f hgg1c.pk012.m2.f hgg1c.pk015.h23 hgg1c.pk006.n2 hgg1c.pk015.h23.f hgg1c.pk006.n2.f | 14 |
| 20G04 | AF500022 | hgg1c.pk004.l24 hgg1c.pk004.l24.f hgg1c.pk012.g22.f hgg1c.pk008.b16.f hgg1c.pk012.g22 hgg1c.pk008.b16 | 4 |
| 22C12 | AF500029 | hgg1c.pk002.d4.f hgg1c.pk005.a4.f hgg1c.pk007.k11.f hgg1c.pk052.f15 hgg1c.pk048.m17.f hgg1c.pk049.m17.f hgg1c.pk013.g17.f hgg1c.pk052.f15.f hgg1c.pk049.m17 hgg1c.pk050.l6.f hgg1c.pk048.m17 hgg1c.pk013.g17 hgg1c.pk050.l6 hgg1c.pk005.a4 hgg1c.pk007.k11 hgg1c.pk002.d4 | 16 |
| 25G01 | AF006052 | hgg1c.pk009.c16.f hgg1c.pk009.c16 | 2 |
| 27D09 | AY101190 | hgg1c.pk012.g22.f hgg1c.pk008.b16.f hgg1c.pk012.g22 hgg1c.pk008.b16 | 4 |
| 28B03 | AF500025 | hgg1c.pk009.i9 hgg1c.pk009.i8 | 2 |
| 29D09 | AF500016 | hgg1c.pk050.k7.f hgg1c.pk014.o19 | 2 |
| 30C02 | AF502393 | hgg1c.pk015.h22.f hgg1c.pk051.b12.f hgg1c.pk015.h22 hgg1c.pk051.b12 | 4 |
| 30E03 | AF500035 | hgg1c.pk051.b7 hgg1c.pk051.b7.f hgg1c.pk006.o5 hgg1c.pk012.n15 hgg1c.pk012.n15.f hgg1c.pk014.e3.f hgg1c.pk006.o5.f | 7 |
| 32E03 | AF500036 | hgg1c.pk013.h5 hgg1c.pk013.h5.f hgg1c.pk001.m24 hgg1c.pk002.l10 hgg1c.pk002.l10.f hgg1c.pk001.m24.f hgg1c.pk014.l23.f hgg1c.pk002.p3.f hgg1c.pk010.k19 hgg1c.pk010.k19.f | 10 |
| 45D07 | AF520565 | hgg1c.pk015.m23.f hgg1c.pk015.m23 | 2 |

# APPENDIX B. CHAPTER 3 SUPPLEMENTARY INFORMATION

## Supplementary Figures

**Fig. S1. Raw Maximum Likelihood phylogenetic tree of the GCN5-related N-acetyltransferase (GNAT) superfamily and newly identified GNATs similar to Hoplolaimina homologs.** Branches for each GNAT family are color-coded accordingly. Bootstrap support values are indicated at corresponding nodes, and those that support monophyly of each GNAT family are oversized in red font. Names and organisms found to contain each GNAT family are indicated to the right with brackets around the respective leaves. Sequence identifications are provided in parentheses at each leaf.

**Fig. S2. Detailed Maximum Likelihood phylogenetic tree of the FAM7 GNATs including Hoplolaimina homologs.** Branches for each phylogenetic group are color-coded according to their taxonomy. Bootstrap support values are indicated at corresponding nodes, and those that support possible horizontal gene transfer (HGT) events are oversized in red font. Organism names are indicated to the right with brackets around the respective leaves. Sequence identifications are provided in parentheses at each leaf.

**Fig. S3. Detailed Maximum Likelihood phylogenetic tree of INVs similar to Hoplolaimina homologs.** Branches for each phylogenetic group are color-coded according to their taxonomy. Bootstrap support values are indicated at corresponding nodes, and those that support possible horizontal gene transfer (HGT) events are oversized in red font. Organism names are indicated to the right with brackets around the respective leaves. Sequence identifications are provided in parentheses at each leaf.

**Fig. S4. Detailed Maximum Likelihood phylogenetic tree of CMs similar to Hoplolaimina homologs.** Branches for each phylogenetic group are color-coded according to their taxonomy. Bootstrap support values are indicated at corresponding nodes, and those that support possible horizontal gene transfer (HGT) events are oversized in red font. Organism names are indicated to the right with brackets around the respective leaves. Sequence identifications are provided in parentheses at each leaf. Note that the *Nacobbus aberrans* leaf is colored red to indicate its possible subfunctionalization from the other Hoplolaimina homologs (see Supplementary Text).

**Fig. S5. Detailed Maximum Likelihood phylogenetic tree of initially poorly clustered FAM7 GNATs.** Branches for each phylogenetic group are color-coded according to their taxonomy. Bootstrap support values are indicated at corresponding nodes, and those that support possible horizontal gene transfer (HGT) events are oversized in red font. Organism names are indicated to the right with brackets around the respective leaves. Sequence identifications are provided in parentheses at each leaf.

**Fig. S6. Sequence alignment of Hoplolaimina GLAND1 (FAM7 GNAT) proteins.** Purple shading illustrates conserved amino acids. A consensus sequence is provided below the alignment. Amino acid positions are indicated to the left and right of each sequence in each row. A green bracket is drawn around the N-terminal signal peptides and a red bracket around the GNAT regions. Notice the extent of conserved amino acids within the GNAT regions, as opposed to the minimal conservation observed outside of the GNAT regions. Hg=*Heterodera glycines*, Ha=*Heterodera avenae*, Gp=*Globodera pallida*, Gr=*Globodera rostochiensis*, *Rr=Rotylenchus reneformis*.

**Fig. S7. Evolution of INVs in Hoplolaimina following horizontal gene transfer (HGT) from rhizobacteria. (A)** Subtree of the cluster containing Hoplolaimina plant parasitic nematodes (PPN) and rhizobacteria (Rhizobiales) calculated from the raw phylogenetic tree in Figure S3. Hoplolaimina PPN branches are colored green and rhizobacteria branches are colored red. Bootstrap support values are indicated at corresponding nodes. Sequence identifications are provided in parentheses at each leaf. Supported clusters are indicated at corresponding nodes with oversized, black bold italicized font. **(B)** Schematic diagrams of all identified INVs in Hoplolaimina PPN, also listed in panel A. Probable N-terminal signal peptides (SP) are illustrated in green. Regions predicted for transmembrane (TM) domains, with or without an overlapping predicted SP, are illustrated in gray. INV domains (GH32; glycoside hydrolase family 32) are illustrated in red. Protein schematics are drawn according to the scale provided at the bottom. Hg=*Heterodera glycines*, Ha=*Heterodera avenae*, Gr=*Globodera rostochiensis*, Gp=*Globodera pallida*, Rr=*Rotylenchus reneformis*, Na=*Nacobbus aberrans*, Mi=*Meloidogyne incognita*.

**Fig. S8. Evolution of CMs in Hoplolaimina following horizontal gene transfer (HGT) from *Burkholderia*-related bacteria. (A)** Subtree of the cluster containing Hoplolaimina plant parasitic nematodes (PPN) and *Burkholderia* spp. calculated from the raw phylogenetic tree in Figure S4. Hoplolaimina PPN branches are colored green and *Burkholderia* branches are colored red. Bootstrap support values are indicated at corresponding nodes. Sequence identifications are provided in parentheses at each leaf. Supported clusters are indicated at corresponding nodes with oversized, black bold italicized font. **(B)** Schematic diagrams of all identified CMs in Hoplolaimina PPN, also listed in panel A. N-terminal signal peptides (SP) are illustrated in green. Transmembrane (TM) regions are illustrated in gray. CM type 2 domains (CM_2s) are illustrated in purple. Proteins with incomplete C-terminal ends are illustrated with dots (…). Protein schematics are drawn according to the scale provided at the bottom. Hoplolaimina PPN groups for all proteins are indicated to the left. Hg=*Heterodera glycines*, Ha=*Heterodera avenae*, Hs=*Heterodera schachtii*, Gr=*Globodera rostochiensis*, Gp=*Globodera pallida*, Gt=*Globodera tabacum*, Ge=*Globodera ellingtonae*, Rr=*Rotylenchus reneformis*, Na=*Nacobbus aberrans*, Mi=*Meloidogyne incognita*, Ma=*Meloidogyne artiellia*, Mj=*Meloidogyne javanica*.

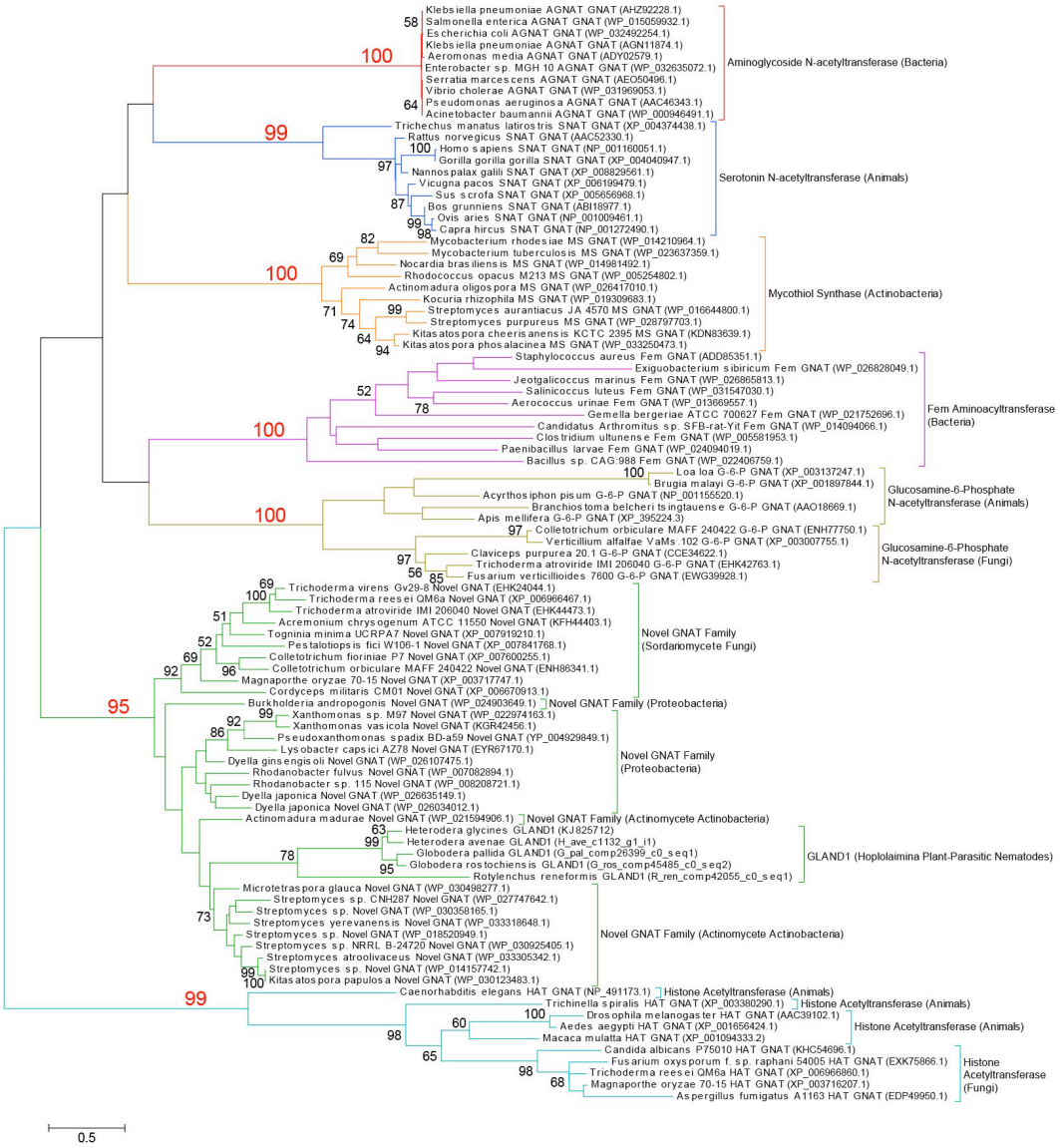

Klebsiella pneumoniae AGNAT GNAT (AHZ92228.1)
Salmonella enterica AGNAT GNAT (WP_015059932.1)
Escherichia coli AGNAT GNAT (WP_032492254.1)
Klebsiella pneumoniae AGNAT GNAT (AGN11874.1)
Aeromonas media AGNAT GNAT (ADY02579.1)
Enterobacter sp. MGH 10 AGNAT GNAT (WP_032635072.1)
Serratia marcescens AGNAT GNAT (AEO50496.1)
Vibrio cholerae AGNAT GNAT (WP_031969053.1)
Pseudomonas aeruginosa AGNAT GNAT (AAC46343.1)
Acinetobacter baumannii AGNAT GNAT (WP_000946491.1)
Aminoglycoside N-acetyltransferase (Bacteria)
Trichechus manatus latirostris SNAT GNAT (XP_004374438.1)
Rattus norvegicus SNAT GNAT (AAC52330.1)
Homo sapiens SNAT GNAT (NP_001160051.1)
Gorilla gorilla gorilla SNAT GNAT (XP_004040947.1)
Nannospalax galili SNAT GNAT (XP_008829561.1)
Vicugna pacos SNAT GNAT (XP_006199479.1)
Sus scrofa SNAT GNAT (XP_005656968.1)
Bos grunniens SNAT GNAT (ABI18977.1)
Ovis aries SNAT GNAT (NP_001009461.1)
Capra hircus SNAT GNAT (NP_001272490.1)
Serotonin N-acetyltransferase (Animals)
Mycobacterium rhodesiae MS GNAT (WP_014210964.1)
Mycobacterium tuberculosis MS GNAT (WP_023637359.1)
Nocardia brasiliensis MS GNAT (WP_014981492.1)
Rhodococcus opacus M213 MS GNAT (WP_005254802.1)
Actinomadura oligospora MS GNAT (WP_026417010.1)
Kocuria rhizophila MS GNAT (WP_019309683.1)
Streptomyces aurantiacus JA 4570 MS GNAT (WP_016644800.1)
Streptomyces purpureus MS GNAT (WP_02879703.1)
Kitasatospora cheerisanensis KCTC 2395 MS GNAT (KDN83639.1)
Kitasatospora phosalacinea MS GNAT (WP_033250473.1)
Mycothiol Synthase (Actinobacteria)
Staphylococcus aureus Fem GNAT (ADD85351.1)
Exiguobacterium sibiricum Fem GNAT (WP_026828049.1)
Jeotgalicoccus marinus Fem GNAT (WP_026865813.1)
Salinicoccus luteus Fem GNAT (WP_031547030.1)
Aerococcus urinae Fem GNAT (WP_013669557.1)
Gemella bergeriae ATCC 700627 Fem GNAT (WP_021752696.1)
Candidatus Arthromitus sp. SFB-rat-Yit Fem GNAT (WP_014094066.1)
Clostridium ultunense Fem GNAT (WP_005581953.1)
Paenibacillus larvae Fem GNAT (WP_024094019.1)
Bacillus sp. CAG:988 Fem GNAT (WP_022406759.1)
Fem Aminoacyltransferase (Bacteria)
Loa loa G-6-P GNAT (XP_003137247.1)
Brugia malayi G-6-P GNAT (XP_001897844.1)
Acyrthosiphon pisum G-6-P GNAT (NP_001155520.1)
Branchiostoma belcheri tsingtauense G-6-P GNAT (AAO18669.1)
Apis mellifera G-6-P GNAT (XP_395224.3)
Glucosamine-6-Phosphate N-acetyltransferase (Animals)
Colletotrichum orbiculare MAFF 240422 G-6-P GNAT (ENH77750.1)
Verticillium alfalfae VaMs.102 G-6-P GNAT (XP_003007755.1)
Claviceps purpurea 20.1 G-6-P GNAT (CCE34622.1)
Trichoderma atroviride IMI 206040 G-6-P GNAT (EHK42763.1)
Fusarium verticillioides 7600 G-6-P GNAT (EWG39928.1)
Glucosamine-6-Phosphate N-acetyltransferase (Fungi)
Trichoderma virens Gv29-8 Novel GNAT (EHK24044.1)
Trichoderma reesei QM6a Novel GNAT (XP_006966467.1)
Trichoderma atroviride IMI 206040 Novel GNAT (EHK44473.1)
Acremonium chrysogenum ATCC 11550 Novel GNAT (KFH44403.1)
Togninia minima UCRPA7 Novel GNAT (XP_007919210.1)
Pestalotiopsis fici W106-1 Novel GNAT (XP_007841768.1)
Colletotrichum fioriniae PJ7 Novel GNAT (XP_007600255.1)
Colletotrichum orbiculare MAFF 240422 Novel GNAT (ENH86341.1)
Magnaporthe oryzae 70-15 Novel GNAT (XP_003717747.1)
Cordyceps militaris CM01 Novel GNAT (XP_006670913.1)
Novel GNAT Family (Sordariomycete Fungi)
Burkholderia andropogonis Novel GNAT (WP_024903649.1)
Novel GNAT Family (Proteobacteria)
Xanthomonas sp. M97 Novel GNAT (WP_022974163.1)
Xanthomonas vasicola Novel GNAT (KGR42456.1)
Pseudoxanthomonas spadix BD-a59 Novel GNAT (YP_004929849.1)
Lysobacter capsici AZ78 Novel GNAT (EYR67170.1)
Dyella ginsengisoli Novel GNAT (WP_026107475.1)
Rhodanobacter fulvus Novel GNAT (WP_007082894.1)
Rhodanobacter sp. 115 Novel GNAT (WP_008208721.1)
Dyella japonica Novel GNAT (WP_026635149.1)
Dyella japonica Novel GNAT (WP_026034012.1)
Novel GNAT Family (Proteobacteria)
Actinomadura madurae Novel GNAT (WP_021594906.1)
Novel GNAT Family (Actinomycete Actinobacteria)
Heterodera glycines GLAND1 (KJ825712)
Heterodera avenae GLAND1 (H_ave_c1132_g1_i1)
Globodera pallida GLAND1 (G_pal_comp26399_c0_seq1)
Globodera rostochiensis GLAND1 (G_ros_comp45485_c0_seq2)
Rotylenchus reneformis GLAND1 (R_ren_comp42055_c0_seq1)
GLAND1 (Hoplolaimina Plant-Parasitic Nematodes)
Microtetraspora glauca Novel GNAT (WP_030498277.1)
Streptomyces sp. CNH287 Novel GNAT (WP_027747642.1)
Streptomyces sp. Novel GNAT (WP_030358165.1)
Streptomyces yerevanensis Novel GNAT (WP_033318648.1)
Streptomyces sp. Novel GNAT (WP_018520949.1)
Streptomyces sp. NRRL B-24720 Novel GNAT (WP_030925405.1)
Streptomyces atroolivaceus Novel GNAT (WP_033305342.1)
Streptomyces sp. Novel GNAT (WP_014157742.1)
Kitasatospora papulosa Novel GNAT (WP_030123483.1)
Novel GNAT Family (Actinomycete Actinobacteria)
Caenorhabditis elegans HAT GNAT (NP_491173.1)
Histone Acetyltransferase (Animals)
Trichinella spiralis HAT GNAT (XP_003380290.1)
Histone Acetyltransferase (Animals)
Drosophila melanogaster HAT GNAT (AAC39102.1)
Aedes aegypti HAT GNAT (XP_001656424.1)
Macaca mulatta HAT GNAT (XP_001094333.2)
Histone Acetyltransferase (Animals)
Candida albicans P75010 HAT GNAT (KHC54696.1)
Fusarium oxysporum f. sp. raphani 54005 HAT GNAT (EXK75866.1)
Trichoderma reesei QM6a HAT GNAT (XP_006966860.1)
Magnaporthe oryzae 70-15 HAT GNAT (XP_003716207.1)
Aspergillus fumigatus A1163 HAT GNAT (EDP49950.1)
Histone Acetyltransferase (Fungi)
0.5

Figure S1

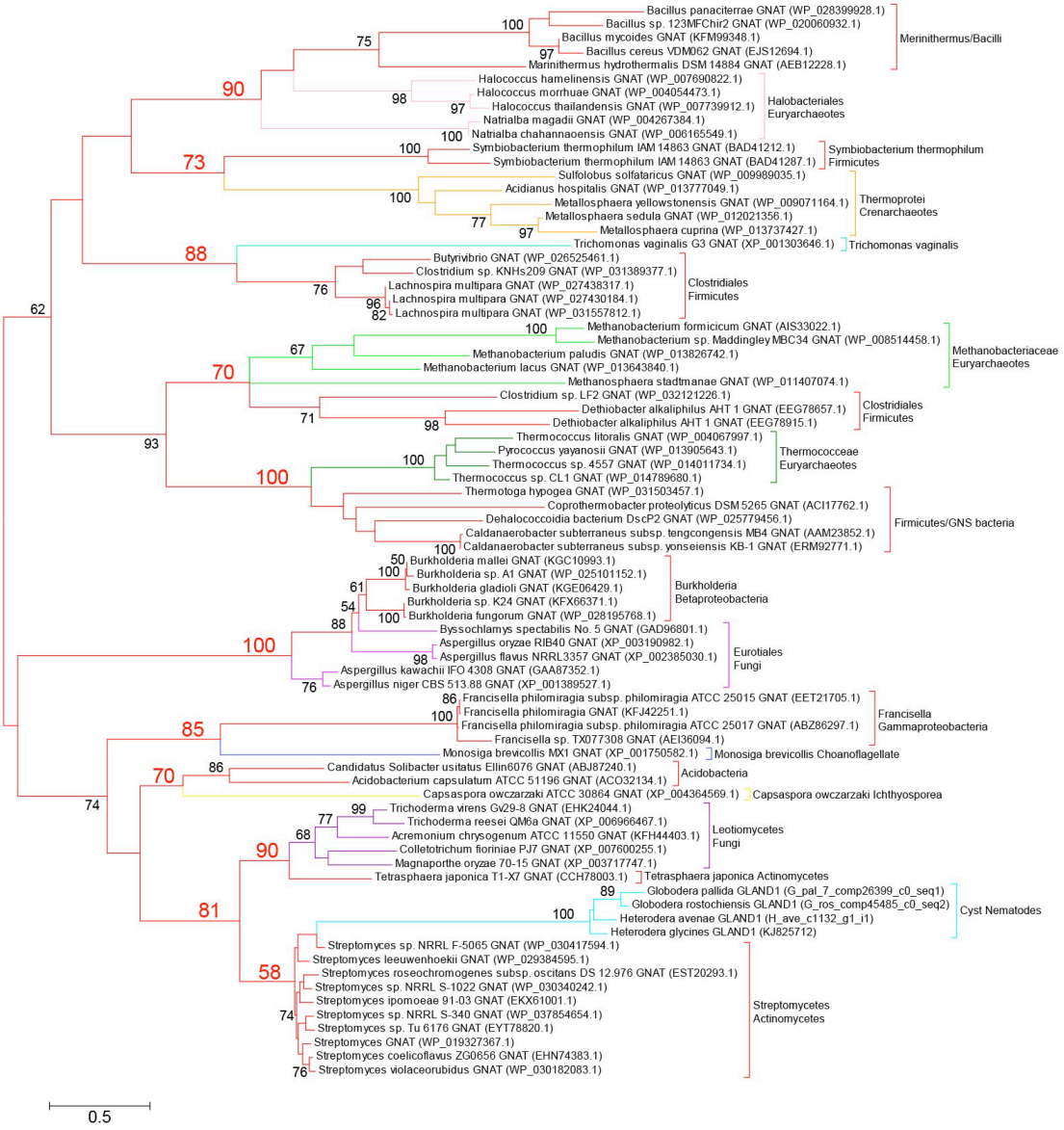

Bacillus panaciterrae GNAT (WP_028399928.1)
Bacillus sp. 123MFChir2 GNAT (WP_020060932.1)
Bacillus mycoides GNAT (KFM99348.1)
Bacillus cereus VDM062 GNAT (EJS12694.1)
Marinithermus hydrothermalis DSM 14884 GNAT (AEB12228.1)
Merinithermus/Bacilli
Halococcus hamelinensis GNAT (WP_007690822.1)
Halococcus morrhuae GNAT (WP_004054473.1)
Halococcus thailandensis GNAT (WP_007739912.1)
Natrialba magadii GNAT (WP_004267384.1)
Natrialba chahannaoensis GNAT (WP_006165549.1)
Halobacteriales
Euryarchaeotes
Symbiobacterium thermophilum IAM 14863 GNAT (BAD41212.1)
Symbiobacterium thermophilum IAM 14863 GNAT (BAD41287.1)
Symbiobacterium thermophilum
Firmicutes
Sulfolobus solfataricus GNAT (WP_009989035.1)
Acidianus hospitalis GNAT (WP_013777049.1)
Metallosphaera yellowstonensis GNAT (WP_009071164.1)
Metallosphaera sedula GNAT (WP_012021356.1)
Metallosphaera cuprina GNAT (WP_013737427.1)
Thermoprotei
Crenarchaeotes
Trichomonas vaginalis G3 GNAT (XP_001303646.1)
Trichomonas vaginalis
Butyrivibrio GNAT (WP_026525461.1)
Clostridium sp. KNHs209 GNAT (WP_031389377.1)
Lachnospira multipara GNAT (WP_027438317.1)
Lachnospira multipara GNAT (WP_027430184.1)
Lachnospira multipara GNAT (WP_031557812.1)
Clostridiales
Firmicutes
Methanobacterium formicicum GNAT (AIS33022.1)
Methanobacterium sp. Maddingley MBC34 GNAT (WP_008514458.1)
Methanobacterium paludis GNAT (WP_013826742.1)
Methanobacterium lacus GNAT (WP_013643840.1)
Methanosphaera stadtmanae GNAT (WP_011407074.1)
Methanobacteriaceae
Euryarchaeotes
Clostridium sp. LF2 GNAT (WP_032121226.1)
Dethiobacter alkaliphilus AHT 1 GNAT (EEG78657.1)
Dethiobacter alkaliphilus AHT 1 GNAT (EEG78915.1)
Clostridiales
Firmicutes
Thermococcus litoralis GNAT (WP_004067997.1)
Pyrococcus yayanosii GNAT (WP_013905643.1)
Thermococcus sp. 4557 GNAT (WP_014011734.1)
Thermococcus sp. CL1 GNAT (WP_014789680.1)
Thermococcaceae
Euryarchaeotes
Thermotoga hypogea GNAT (WP_031503457.1)
Coprothermobacter proteolyticus DSM 5265 GNAT (ACI17762.1)
Dehalococcoidia bacterium DscP2 GNAT (WP_025779456.1)
Caldanaerobacter subterraneus subsp. tengcongensis MB4 GNAT (AAM23852.1)
Caldanaerobacter subterraneus subsp. yonseiensis KB-1 GNAT (ERM92771.1)
Firmicutes/GNS bacteria
Burkholderia mallei GNAT (KGC10993.1)
Burkholderia sp. A1 GNAT (WP_025101152.1)
Burkholderia gladioli GNAT (KGE06429.1)
Burkholderia sp. K24 GNAT (KFX66371.1)
Burkholderia fungorum GNAT (WP_028195768.1)
Burkholderia
Betaproteobacteria
Byssochlamys spectabilis No. 5 GNAT (GAD96801.1)
Aspergillus oryzae RIB40 GNAT (XP_003190982.1)
Aspergillus flavus NRRL3357 GNAT (XP_002385030.1)
Aspergillus kawachii IFO 4308 GNAT (GAA87352.1)
Aspergillus niger CBS 513.88 GNAT (XP_001389527.1)
Eurotiales
Fungi
Francisella philomiragia subsp. philomiragia ATCC 25015 GNAT (EET21705.1)
Francisella philomiragia GNAT (KFJ42251.1)
Francisella philomiragia subsp. philomiragia ATCC 25017 GNAT (ABZ86297.1)
Francisella sp. TX077308 GNAT (AEI36094.1)
Francisella
Gammaproteobacteria
Monosiga brevicollis MX1 GNAT (XP_001750582.1)
Monosiga brevicollis Choanoflagellate
Candidatus Solibacter usitatus Ellin6076 GNAT (ABJ87240.1)
Acidobacterium capsulatum ATCC 51196 GNAT (ACO32134.1)
Acidobacteria
Capsaspora owczarzaki ATCC 30864 GNAT (XP_004364569.1)
Capsaspora owczarzaki Ichthyosporea
Trichoderma virens Gv29-8 GNAT (EHK24044.1)
Trichoderma reesei QM6a GNAT (XP_006966467.1)
Acremonium chrysogenum ATCC 11550 GNAT (KFH44403.1)
Colletotrichum fioriniae PJ7 GNAT (XP_007600255.1)
Magnaporthe oryzae 70-15 GNAT (XP_003717747.1)
Leotiomycetes
Fungi
Tetrasphaera japonica T1-X7 GNAT (CCH78003.1)
Tetrasphaera japonica Actinomycetes
Globodera pallida GLAND1 (G_pal_7_comp26399_c0_seq1)
Globodera rostochiensis GLAND1 (G_ros_comp45485_c0_seq2)
Heterodera avenae GLAND1 (H_ave_c1132_g1_i1)
Heterodera glycines GLAND1 (KJ825712)
Cyst Nematodes
Streptomyces sp. NRRL F-5065 GNAT (WP_030417594.1)
Streptomyces leeuwenhoekii GNAT (WP_029384595.1)
Streptomyces roseochromogenes subsp. oscitans DS 12.976 GNAT (EST20293.1)
Streptomyces sp. NRRL S-1022 GNAT (WP_030340242.1)
Streptomyces ipomoeae 91-03 GNAT (EKX61001.1)
Streptomyces sp. NRRL S-340 GNAT (WP_037854654.1)
Streptomyces sp. Tu 6176 GNAT (EYT78820.1)
Streptomyces GNAT (WP_019327367.1)
Streptomyces coelicoflavus ZG0656 GNAT (EHN74383.1)
Streptomyces violaceorubidus GNAT (WP_030182083.1)
Streptomycetes
Actinomycetes
0.5

Figure S2

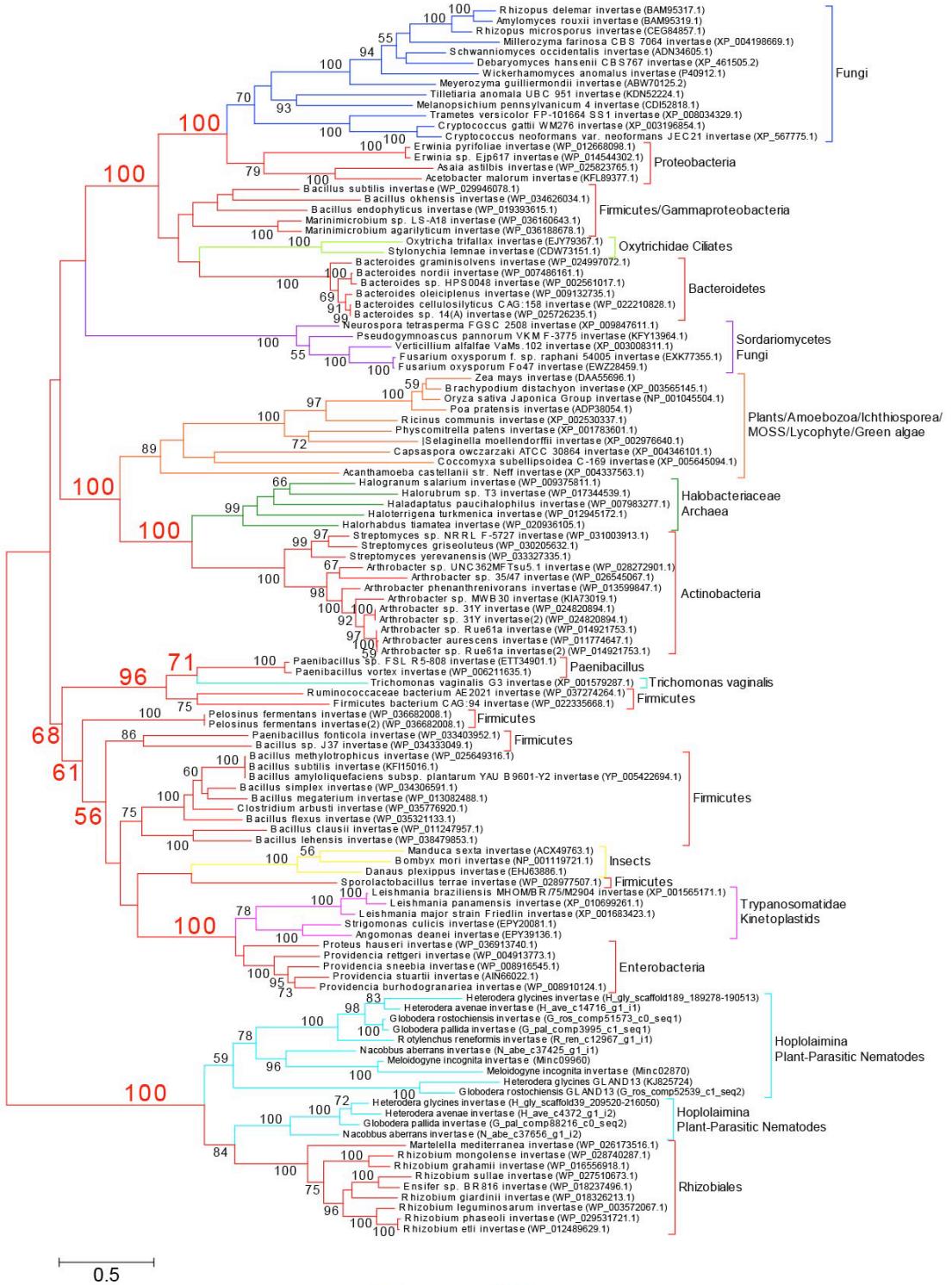

Rhizopus delemar invertase (BAM95317.1)
Amylomyces rouxii invertase (BAM95319.1)
Rhizopus microsporus invertase (CEG84857.1)
Millerozyma farinosa CBS 7064 invertase (XP_004198669.1)
Schwanniomyces occidentalis invertase (ADN34605.1)
Debaryomyces hansenii CBS767 invertase (XP_461505.2)
Wickerhamomyces anomalus invertase (P40912.1)
Meyerozyma guilliermondii invertase (ABW70125.2)
Tilletiaria anomala UBC 951 invertase (KDN52224.1)
Melanopsichium pennsylvanicum 4 invertase (CDI52818.1)
Trametes versicolor FP-101664 SS1 invertase (XP_008034329.1)
Cryptococcus gattii WM276 invertase (XP_003196854.1)
Cryptococcus neoformans var. neoformans JEC21 invertase (XP_567775.1)
Fungi
Erwinia pyrifoliae invertase (WP_012668098.1)
Erwinia sp. Ejp617 invertase (WP_014544302.1)
Asaia astilbis invertase (WP_025823765.1)
Acetobacter malorum invertase (KFL89377.1)
Proteobacteria
Bacillus subtilis invertase (WP_029946078.1)
Bacillus okhensis invertase (WP_034626034.1)
Bacillus endophyticus invertase (WP_019393615.1)
Marinimicrobium sp. LS-A18 invertase (WP_036160643.1)
Marinimicrobium agarilyticum invertase (WP_036188678.1)
Firmicutes/Gammaproteobacteria
Oxytricha trifallax invertase (EJY79367.1)
Stylonychia lemnae invertase (CDW73151.1)
Oxytrichidae Ciliates
Bacteroides graminisolvens invertase (WP_024997072.1)
Bacteroides nordii invertase (WP_007486161.1)
Bacteroides sp. HPS0048 invertase (WP_002561017.1)
Bacteroides oleiciplenus invertase (WP_009132735.1)
Bacteroides cellulosilyticus CAG:158 invertase (WP_022210828.1)
Bacteroides sp. 14(A) invertase (WP_025726235.1)
Bacteroidetes
Neurospora tetrasperma FGSC 2508 invertase (XP_009847611.1)
Pseudogymnoascus pannorum VKMF-3775 invertase (KFY13964.1)
Verticillium alfalfae VaMs.102 invertase (XP_003008311.1)
Fusarium oxysporum f. sp. raphani 54005 invertase (EXK77355.1)
Fusarium oxysporum Fo47 invertase (EWZ28459.1)
Sordariomycetes
Fungi
Zea mays invertase (DAA55696.1)
Brachypodium distachyon invertase (XP_003565145.1)
Oryza sativa Japonica Group invertase (NP_001045504.1)
Poa pratensis invertase (ADP38054.1)
Ricinus communis invertase (XP_002530337.1)
Physcomitrella patens invertase (XP_001783601.1)
Selaginella moellendorffii invertase (XP_002976640.1)
Capsaspora owczarzaki ATCC 30864 invertase (XP_004346101.1)
Coccomyxa subellipsoidea C-169 invertase (XP_005645094.1)
Acanthamoeba castellanii str. Neff invertase (XP_004337563.1)
Plants/Amoebozoa/Ichthiosporea/
MOSS/Lycophyte/Green algae
Halogranum salarium invertase (WP_009375811.1)
Halorubrum sp. T3 invertase (WP_017344539.1)
Haladaptatus paucihalophilus invertase (WP_007983277.1)
Haloterrigena turkmenica invertase (WP_012945172.1)
Halorhabdus tiamatea invertase (WP_020936105.1)
Halobacteriaceae
Archaea
Streptomyces sp. NRRL F-5727 invertase (WP_031003913.1)
Streptomyces griseoluteus (WP_030205632.1)
Streptomyces yerevanensis (WP_033327335.1)
Arthrobacter sp. UNC362MFTsu5.1 invertase (WP_028272901.1)
Arthrobacter sp. 35/47 invertase (WP_026545067.1)
Arthrobacter phenanthrenivorans invertase (WP_013599847.1)
Arthrobacter sp. MWB30 invertase (KIA73019.1)
Arthrobacter sp. 31Y invertase (WP_024820894.1)
Arthrobacter sp. 31Y invertase(2) (WP_024820894.1)
Arthrobacter sp. Rue61a invertase (WP_014921753.1)
Arthrobacter aurescens invertase (WP_011774647.1)
Arthrobacter sp. Rue61a invertase(2) (WP_014921753.1)
Actinobacteria
Paenibacillus sp. FSL R5-808 invertase (ETT34901.1)
Paenibacillus vortex invertase (WP_006211635.1)
Paenibacillus
Trichomonas vaginalis G3 invertase (XP_001579287.1)
Trichomonas vaginalis
Ruminococcaceae bacterium AE2021 invertase (WP_037274264.1)
Firmicutes bacterium CAG:94 invertase (WP_022335668.1)
Firmicutes
Pelosinus fermentans invertase (WP_036682008.1)
Pelosinus fermentans invertase(2) (WP_036682008.1)
Firmicutes
Paenibacillus fonticola invertase (WP_033403952.1)
Bacillus sp. J37 invertase (WP_034333049.1)
Firmicutes
Bacillus methylotrophicus invertase (WP_025649316.1)
Bacillus subtilis invertase (KFI15016.1)
Bacillus amyloliquefaciens subsp. plantarum YAU B9601-Y2 invertase (YP_005422694.1)
Bacillus simplex invertase (WP_034306591.1)
Bacillus megaterium invertase (WP_013082488.1)
Clostridium arbusti invertase (WP_035776920.1)
Bacillus flexus invertase (WP_035321133.1)
Bacillus clausii invertase (WP_011247957.1)
Bacillus lehensis invertase (WP_038479853.1)
Firmicutes
Manduca sexta invertase (ACX49763.1)
Bombyx mori invertase (NP_001119721.1)
Danaus plexippus invertase (EHJ63886.1)
Insects
Sporolactobacillus terrae invertase (WP_028977507.1)
Firmicutes
Leishmania braziliensis MHOM/BR/75/M2904 invertase (XP_001565171.1)
Leishmania panamensis invertase (XP_010699261.1)
Leishmania major strain Friedlin invertase (XP_001683423.1)
Strigomonas culicis invertase (EPY20081.1)
Angomonas deanei invertase (EPY39136.1)
Trypanosomatidae
Kinetoplastids
Proteus hauseri invertase (WP_036913740.1)
Providencia rettgeri invertase (WP_004913773.1)
Providencia sneebia invertase (WP_008916545.1)
Providencia stuartii invertase (AIN66022.1)
Providencia burhodogranariea invertase (WP_008910124.1)
Enterobacteria
Heterodera glycines invertase (H_gly_scaffold189_189278-190513)
Heterodera avenae invertase (H_ave_c14716_g1_i1)
Globodera rostochiensis invertase (G_ros_comp51573_c0_seq1)
Globodera pallida invertase (G_pal_comp3995_c1_seq1)
Rotylenchus reneformis invertase (R_ren_c12967_g1_i1)
Nacobbus aberrans invertase (N_abe_c37425_g1_i1)
Meloidogyne incognita invertase (Minc09960)
Meloidogyne incognita invertase (Minc02870)
Heterodera glycines GLAND13 (KJ825724)
Globodera rostochiensis GLAND13 (G_ros_comp52539_c1_seq2)
Hoplolaimina
Plant-Parasitic Nematodes
Heterodera glycines invertase (H_gly_scaffold39_209520-216050)
Heterodera avenae invertase (H_ave_c4372_g1_i2)
Globodera pallida invertase (G_pal_comp88216_c0_seq2)
Nacobbus aberrans invertase (N_abe_c37656_g1_i2)
Hoplolaimina
Plant-Parasitic Nematodes
Martelella mediterranea invertase (WP_026173516.1)
Rhizobium mongolense invertase (WP_028740287.1)
Rhizobium grahamii invertase (WP_016556918.1)
Rhizobium sullae invertase (WP_027510673.1)
Ensifer sp. BR816 invertase (WP_018237496.1)
Rhizobium giardinii invertase (WP_018326213.1)
Rhizobium leguminosarum invertase (WP_003572067.1)
Rhizobium phaseoli invertase (WP_029531721.1)
Rhizobium etli invertase (WP_012489629.1)
Rhizobiales
0.5

Figure S3

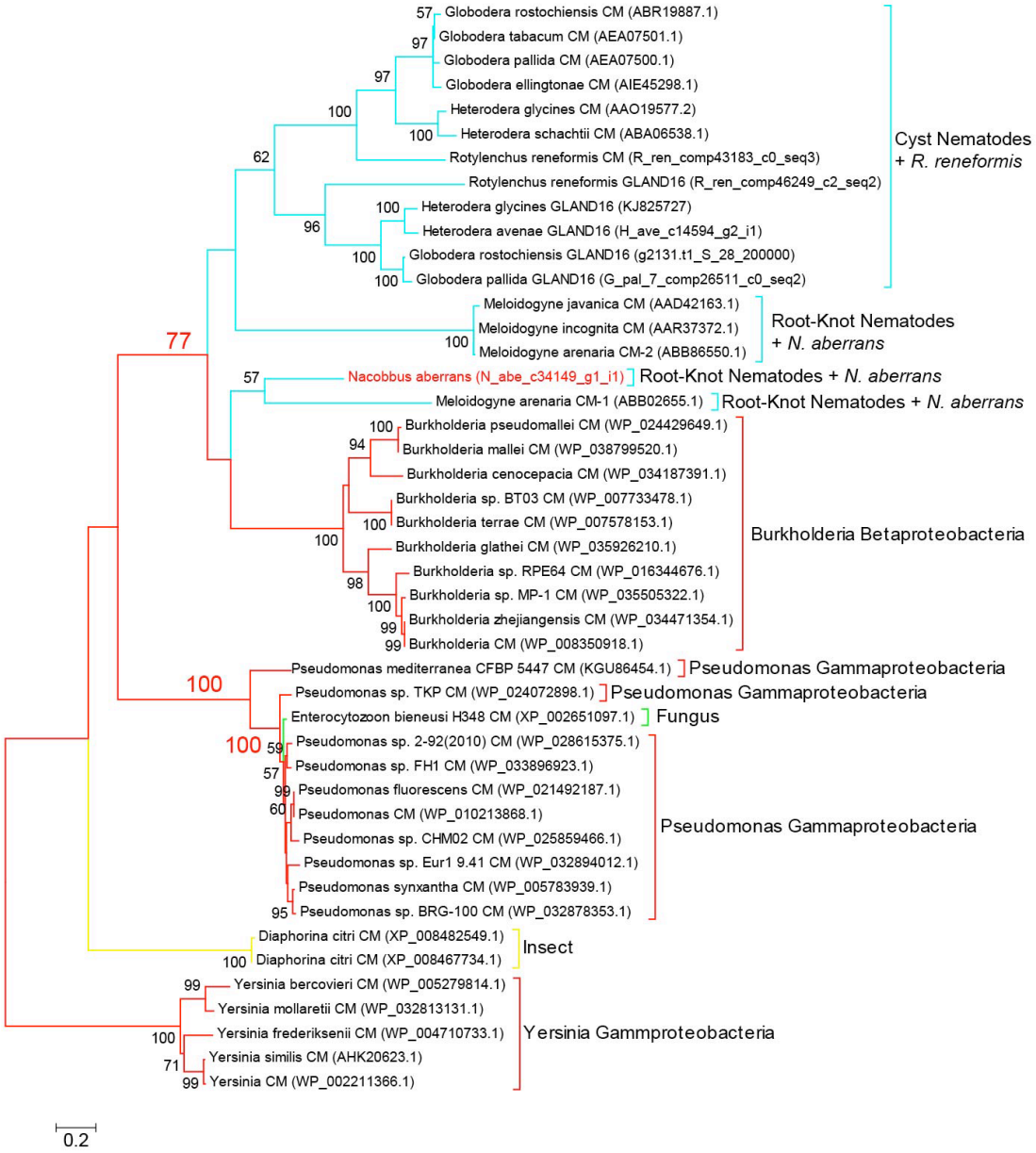

Globodera rostochiensis CM (ABR19887.1)
Globodera tabacum CM (AEA07501.1)
Globodera pallida CM (AEA07500.1)
Globodera ellingtonae CM (AIE45298.1)
Heterodera glycines CM (AAO19577.2)
Heterodera schachtii CM (ABA06538.1)
Rotylenchus reneformis CM (R_ren_comp43183_c0_seq3)
Rotylenchus reneformis GLAND16 (R_ren_comp46249_c2_seq2)
Heterodera glycines GLAND16 (KJ825727)
Heterodera avenae GLAND16 (H_ave_c14594_g2_i1)
Globodera rostochiensis GLAND16 (g2131.t1_S_28_200000)
Globodera pallida GLAND16 (G_pal_7_comp26511_c0_seq2)
Cyst Nematodes + R. reneformis
Meloidogyne javanica CM (AAD42163.1)
Meloidogyne incognita CM (AAR37372.1)
Meloidogyne arenaria CM-2 (ABB86550.1)
Root-Knot Nematodes + N. aberrans
Nacobbus aberrans (N_abe_c34149_g1_i1)
Root-Knot Nematodes + N. aberrans
Meloidogyne arenaria CM-1 (ABB02655.1)
Root-Knot Nematodes + N. aberrans
Burkholderia pseudomallei CM (WP_024429649.1)
Burkholderia mallei CM (WP_038799520.1)
Burkholderia cenocepacia CM (WP_034187391.1)
Burkholderia sp. BT03 CM (WP_007733478.1)
Burkholderia terrae CM (WP_007578153.1)
Burkholderia glathei CM (WP_035926210.1)
Burkholderia sp. RPE64 CM (WP_016344676.1)
Burkholderia sp. MP-1 CM (WP_035505322.1)
Burkholderia zhejiangensis CM (WP_034471354.1)
Burkholderia CM (WP_008350918.1)
Burkholderia Betaproteobacteria
Pseudomonas mediterranea CFBP 5447 CM (KGU86454.1)
Pseudomonas Gammaproteobacteria
Pseudomonas sp. TKP CM (WP_024072898.1)
Pseudomonas Gammaproteobacteria
Enterocytozoon bieneusi H348 CM (XP_002651097.1)
Fungus
Pseudomonas sp. 2-92(2010) CM (WP_028615375.1)
Pseudomonas sp. FH1 CM (WP_033896923.1)
Pseudomonas fluorescens CM (WP_021492187.1)
Pseudomonas CM (WP_010213868.1)
Pseudomonas sp. CHM02 CM (WP_025859466.1)
Pseudomonas sp. Eur1 9.41 CM (WP_032894012.1)
Pseudomonas synxantha CM (WP_005783939.1)
Pseudomonas sp. BRG-100 CM (WP_032878353.1)
Pseudomonas Gammaproteobacteria
Diaphorina citri CM (XP_008482549.1)
Diaphorina citri CM (XP_008467734.1)
Insect
Yersinia bercovieri CM (WP_005279814.1)
Yersinia mollaretii CM (WP_032813131.1)
Yersinia frederiksenii CM (WP_004710733.1)
Yersinia similis CM (AHK20623.1)
Yersinia CM (WP_002211366.1)
Yersinia Gammproteobacteria
0.2

Figure S4

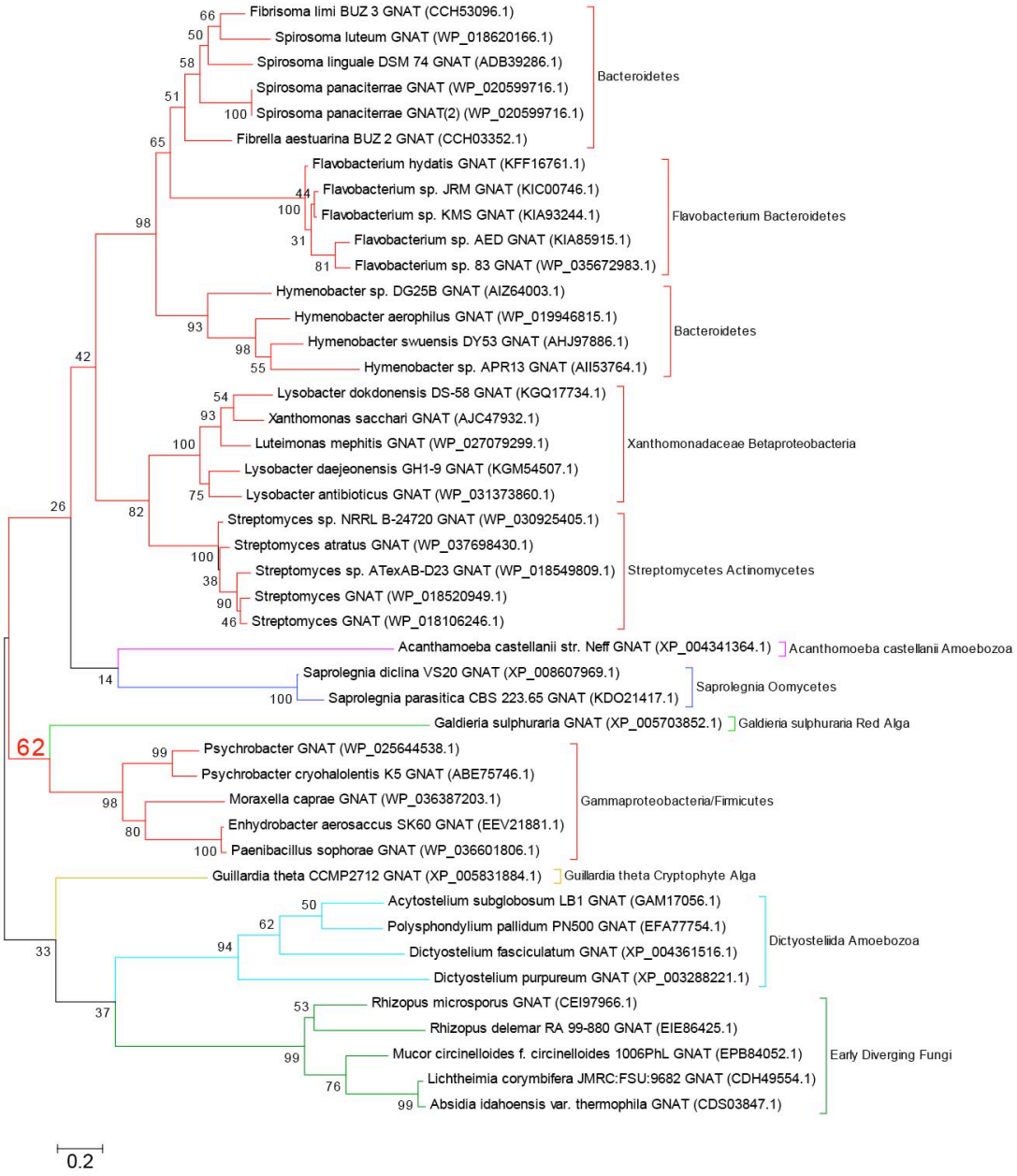

Fibrisoma limi BUZ 3 GNAT (CCH53096.1)
Spirosoma luteum GNAT (WP_018620166.1)
Spirosoma linguale DSM 74 GNAT (ADB39286.1)
Spirosoma panaciterrae GNAT (WP_020599716.1)
Spirosoma panaciterrae GNAT(2) (WP_020599716.1)
Fibrella aestuarina BUZ 2 GNAT (CCH03352.1)
Bacteroidetes
Flavobacterium hydatis GNAT (KFF16761.1)
Flavobacterium sp. JRM GNAT (KIC00746.1)
Flavobacterium sp. KMS GNAT (KIA93244.1)
Flavobacterium sp. AED GNAT (KIA85915.1)
Flavobacterium sp. 83 GNAT (WP_035672983.1)
Flavobacterium Bacteroidetes
Hymenobacter sp. DG25B GNAT (AIZ64003.1)
Hymenobacter aerophilus GNAT (WP_019946815.1)
Hymenobacter swuensis DY53 GNAT (AHJ97886.1)
Hymenobacter sp. APR13 GNAT (AII53764.1)
Bacteroidetes
Lysobacter dokdonensis DS-58 GNAT (KGQ17734.1)
Xanthomonas sacchari GNAT (AJC47932.1)
Luteimonas mephitis GNAT (WP_027079299.1)
Lysobacter daejeonensis GH1-9 GNAT (KGM54507.1)
Lysobacter antibioticus GNAT (WP_031373860.1)
Xanthomonadaceae Betaproteobacteria
Streptomyces sp. NRRL B-24720 GNAT (WP_030925405.1)
Streptomyces atratus GNAT (WP_037698430.1)
Streptomyces sp. ATexAB-D23 GNAT (WP_018549809.1)
Streptomyces GNAT (WP_018520949.1)
Streptomyces GNAT (WP_018106246.1)
Streptomycetes Actinomycetes
Acanthamoeba castellanii str. Neff GNAT (XP_004341364.1)
Acanthomoeba castellanii Amoebozoa
Saprolegnia diclina VS20 GNAT (XP_008607969.1)
Saprolegnia parasitica CBS 223.65 GNAT (KDO21417.1)
Saprolegnia Oomycetes
Galdieria sulphuraria GNAT (XP_005703852.1)
Galdieria sulphuraria Red Alga
Psychrobacter GNAT (WP_025644538.1)
Psychrobacter cryohalolentis K5 GNAT (ABE75746.1)
Moraxella caprae GNAT (WP_036387203.1)
Enhydrobacter aerosaccus SK60 GNAT (EEV21881.1)
Paenibacillus sophorae GNAT (WP_036601806.1)
Gammaproteobacteria/Firmicutes
Guillardia theta CCMP2712 GNAT (XP_005831884.1)
Guillardia theta Cryptophyte Alga
Acytostelium subglobosum LB1 GNAT (GAM17056.1)
Polysphondylium pallidum PN500 GNAT (EFA77754.1)
Dictyostelium fasciculatum GNAT (XP_004361516.1)
Dictyostelium purpureum GNAT (XP_003288221.1)
Dictyostelida Amoebozoa
Rhizopus microsporus GNAT (CEI97966.1)
Rhizopus delemar RA 99-880 GNAT (EIE86425.1)
Mucor circinelloides f. circinelloides 1006PhL GNAT (EPB84052.1)
Lichtheimia corymbifera JMRC:FSU:9682 GNAT (CDH49554.1)
Absidia idahoensis var. thermophila GNAT (CDS03847.1)
Early Diverging Fungi
0.2

Figure S5

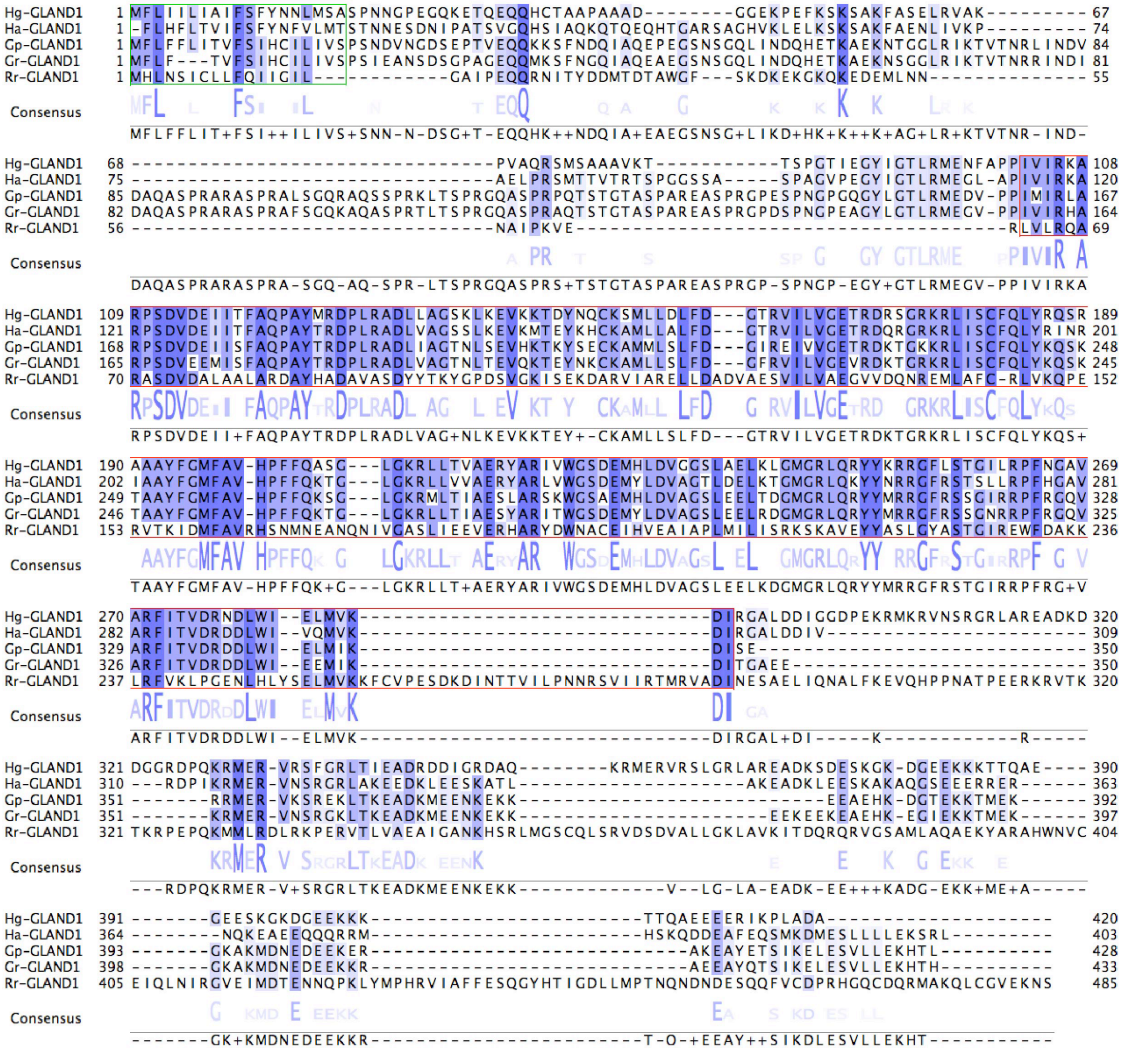



Figure S6

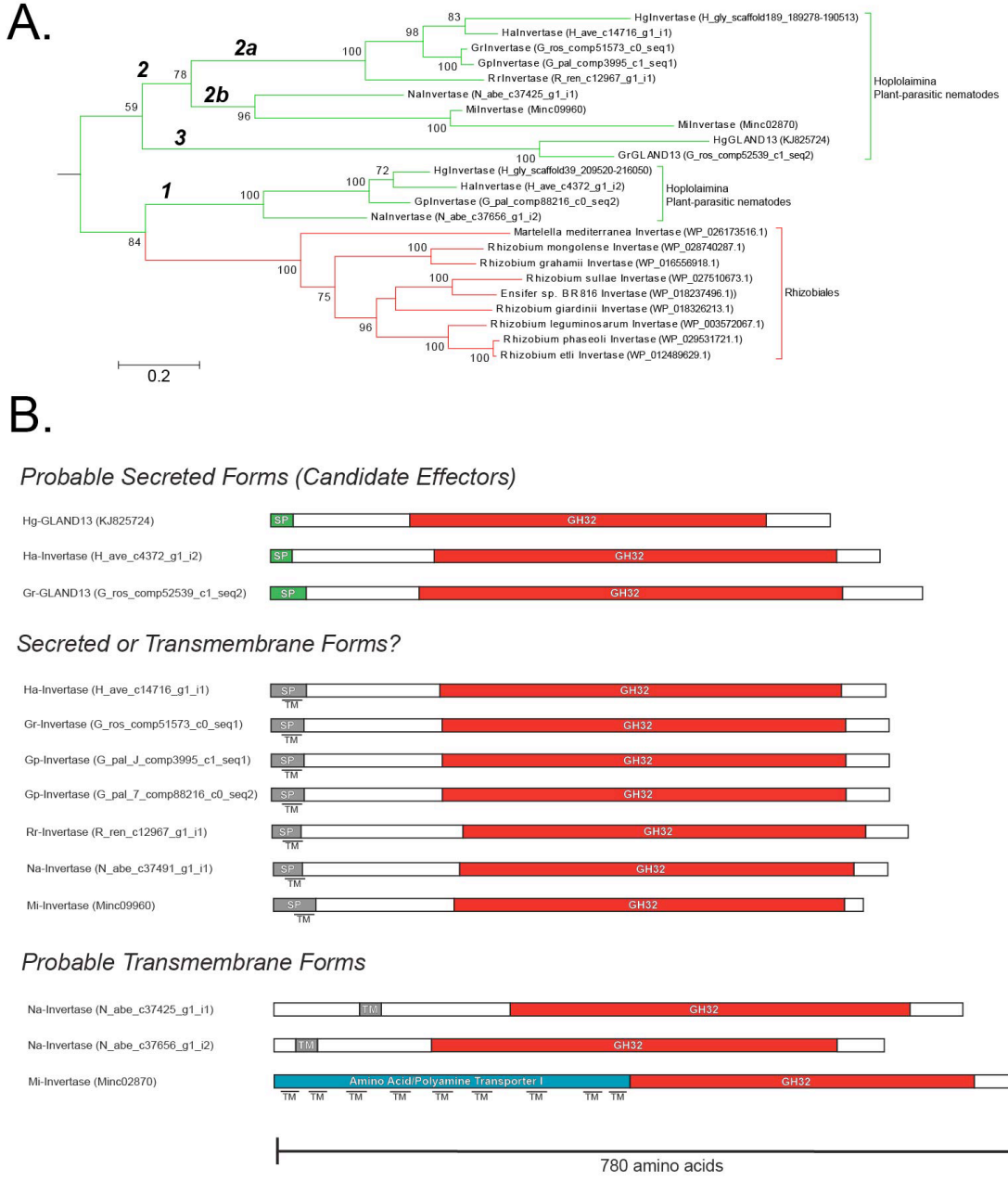

A.
2
2a
2b
3
1
HgInvertase (H_gly_scaffold189_189278-190513)
HaInvertase (H_ave_c14716_g1_i1)
GrInvertase (G_ros_comp51573_c0_seq1)
GpInvertase (G_pal_comp3995_c1_seq1)
RrInvertase (R_ren_c12967_g1_i1)
NaInvertase (N_abe_c37425_g1_i1)
MiInvertase (Minc09960)
MiInvertase (Minc02870)
HgGLAND13 (KJ825724)
GrGLAND13 (G_ros_comp52539_c1_seq2)
Hoplolaimina
Plant-parasitic nematodes
HgInvertase (H_gly_scaffold39_209520-216050)
HaInvertase (H_ave_c4372_g1_i2)
GpInvertase (G_pal_comp88216_c0_seq2)
NaInvertase (N_abe_c37656_g1_i2)
Hoplolaimina
Plant-parasitic nematodes
Martelella mediterranea Invertase (WP_026173516.1)
Rhizobium mongolense Invertase (WP_028740287.1)
Rhizobium grahamii Invertase (WP_016556918.1)
Rhizobium sullae Invertase (WP_027510673.1)
Ensifer sp. BR816 Invertase (WP_018237496.1))
Rhizobium giardinii Invertase (WP_018326213.1)
Rhizobium leguminosarum Invertase (WP_003572067.1)
Rhizobium phaseoli Invertase (WP_029531721.1)
Rhizobium etli Invertase (WP_012489629.1)
Rhizobiales
0.2
B.
Probable Secreted Forms (Candidate Effectors)
Hg-GLAND13 (KJ825724)
Ha-Invertase (H_ave_c4372_g1_i2)
Gr-GLAND13 (G_ros_comp52539_c1_seq2)
SP
GH32
Secreted or Transmembrane Forms?
Ha-Invertase (H_ave_c14716_g1_i1)
Gr-Invertase (G_ros_comp51573_c0_seq1)
Gp-Invertase (G_pal_J_comp3995_c1_seq1)
Gp-Invertase (G_pal_7_comp88216_c0_seq2)
Rr-Invertase (R_ren_c12967_g1_i1)
Na-Invertase (N_abe_c37491_g1_i1)
Mi-Invertase (Minc09960)
SP
TM
GH32
Probable Transmembrane Forms
Na-Invertase (N_abe_c37425_g1_i1)
Na-Invertase (N_abe_c37656_g1_i2)
Mi-Invertase (Minc02870)
TM
GH32
Amino Acid/Polyamine Transporter I
780 amino acids

Figure S7

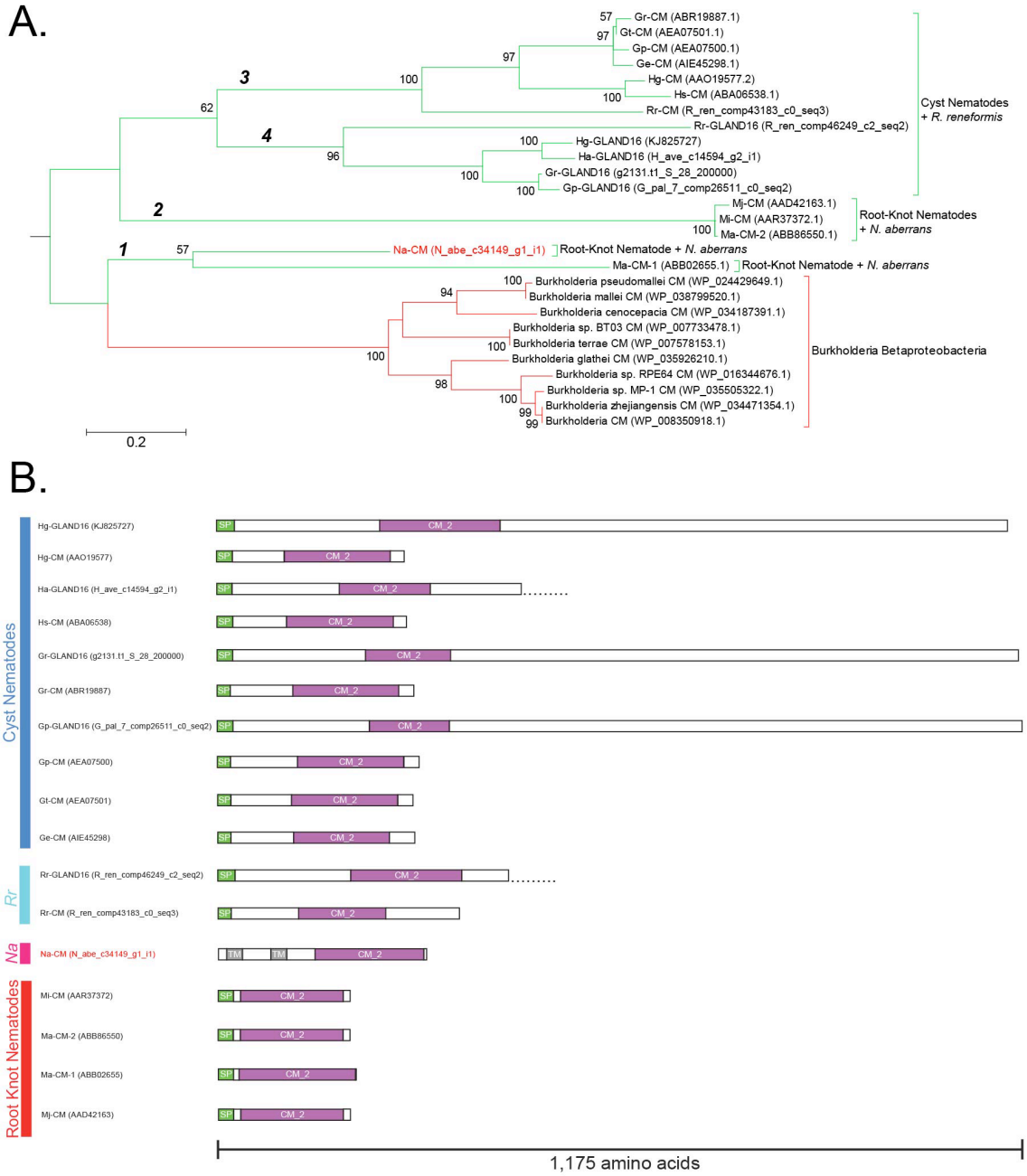

A.
Gr-CM (ABR19887.1)
Gt-CM (AEA07501.1)
Gp-CM (AEA07500.1)
Ge-CM (AIE45298.1)
Hg-CM (AAO19577.2)
Hs-CM (ABA06538.1)
Rr-CM (R_ren_comp43183_c0_seq3)
Rr-GLAND16 (R_ren_comp46249_c2_seq2)
Hg-GLAND16 (KJ825727)
Ha-GLAND16 (H_ave_c14594_g2_i1)
Gr-GLAND16 (g2131.t1_S_28_200000)
Gp-GLAND16 (G_pal_7_comp26511_c0_seq2)
Cyst Nematodes + R. reneformis
Mj-CM (AAD42163.1)
Mi-CM (AAR37372.1)
Ma-CM-2 (ABB86550.1)
Root-Knot Nematodes + N. aberrans
Na-CM (N_abe_c34149_g1_i1)
Root-Knot Nematode + N. aberrans
Ma-CM-1 (ABB02655.1)
Root-Knot Nematode + N. aberrans
Burkholderia pseudomallei CM (WP_024429649.1)
Burkholderia mallei CM (WP_038799520.1)
Burkholderia cenocepacia CM (WP_034187391.1)
Burkholderia sp. BT03 CM (WP_007733478.1)
Burkholderia terrae CM (WP_007578153.1)
Burkholderia glathei CM (WP_035926210.1)
Burkholderia sp. RPE64 CM (WP_016344676.1)
Burkholderia sp. MP-1 CM (WP_035505322.1)
Burkholderia zhejiangensis CM (WP_034471354.1)
Burkholderia CM (WP_008350918.1)
Burkholderia Betaproteobacteria
0.2
B.
Cyst Nematodes
Rr
Na
Root Knot Nematodes
SP
CM_2
TM
1,175 amino acids

Figure S8

## Supplementary Tables

Note that Tables S1, S2 and S4-S9 will be available online at the journal's website upon acceptance for publication. These tables present raw data that were determined to be inessential to the presentation of results for this dissertation. Table S3 presents data pertinent to the presentation of results for this dissertation and is provided below.

**Table S3. Codon adaptiation index analyses of Hoplolaimina *GNAT*, *INV*, and *CM domains* compared to codon usage tables of both Hoplolaimina and donor bacteria**

| accession/ID | organism | codon usage table | result | CAI/ E-CAI | CAI | E-CAI – standard | E-CAI - eubacterial |
|---|---|---|---|---|---|---|---|
| | | *GNAT domains* | | | | | |
| KJ825712 | *Heterodera glycines* | Heterodera glycines [gbinv] | different | 0.983 | 0.744 | 0.757 | n/a |
| H_ave_c1132_g1_i1 | *Heterodera avenae* | Heterodera glycines [gbinv] | similar | 1.028 | 0.760 | 0.739 | n/a |
| G_pal_comp26399_c0_seq1 | *Globodera pallida* | Heterodera glycines [gbinv] | somewhat similar | 1.012 | 0.785 | 0.776 | n/a |
| G_ros_comp45485_c0_seq2 | *Globodera rostochiensis* | Heterodera glycines [gbinv] | somewhat similar | 1.013 | 0.798 | 0.788 | n/a |
| R_ren_comp42055_c0_seq1 | *Rotylenchus reneformis* | Heterodera glycines [gbinv] | somewhat similar | 0.997 | 0.763 | 0.765 | n/a |
| KJ825712 | *Heterodera glycines* | Streptomyces coelicolor A3(2) [gbbct] | different | 0.795 | 0.271 | 0.346 | 0.336 |
| H_ave_c1132_g1_i1 | *Heterodera avenae* | Streptomyces coelicolor A3(2) [gbbct] | different | 0.819 | 0.219 | 0.267 | 0.268 |
| G_pal_comp26399_c0_seq1 | *Globodera pallida* | Streptomyces coelicolor A3(2) [gbbct] | different | 0.829 | 0.294 | 0.36 | 0.349 |
| G_ros_comp45485_c0_seq2 | *Globodera rostochiensis* | Streptomyces coelicolor A3(2) [gbbct] | different | 0.815 | 0.316 | 0.387 | 0.388 |
| R_ren_comp42055_c0_seq1 | *Rotylenchus reneformis* | Streptomyces coelicolor A3(2) [gbbct] | different | 0.843 | 0.228 | 0.273 | 0.268 |
| KJ825712 | *Heterodera glycines* | Streptomyces avermitilis MA-4680 [gbbct] | different | 0.812 | 0.315 | 0.387 | 0.389 |
| H_ave_c1132_g1_i1 | *Heterodera avenae* | Streptomyces avermitilis MA-4680 [gbbct] | different | 0.840 | 0.258 | 0.305 | 0.309 |
| G_pal_comp26399_c0_seq1 | *Globodera pallida* | Streptomyces avermitilis MA-4680 [gbbct] | different | 0.845 | 0.338 | 0.403 | 0.397 |
| G_ros_comp45485_c0_seq2 | *Globodera rostochiensis* | Streptomyces avermitilis MA-4680 [gbbct] | different | 0.834 | 0.362 | 0.433 | 0.435 |
| R_ren_comp42055_c0_seq1 | *Rotylenchus reneformis* | Streptomyces avermitilis MA-4680 [gbbct] | different | 0.853 | 0.270 | 0.314 | 0.319 |
| KJ825712 | *Heterodera glycines* | Streptomyces ambofaciens ATCC 23877 [gbbct] | different | 0.809 | 0.288 | 0.356 | 0.356 |
| H_ave_c1132_g1_i1 | *Heterodera avenae* | Streptomyces ambofaciens ATCC 23877 [gbbct] | different | 0.833 | 0.240 | 0.288 | 0.288 |
| G_pal_comp26399_c0_seq1 | *Globodera pallida* | Streptomyces ambofaciens ATCC 23877 [gbbct] | different | 0.835 | 0.316 | 0.376 | 0.381 |
| G_ros_comp45485_c0_seq2 | *Globodera rostochiensis* | Streptomyces ambofaciens ATCC 23877 [gbbct] | different | 0.825 | 0.338 | 0.412 | 0.407 |
| R_ren_comp42055_c0_seq1 | *Rotylenchus reneformis* | Streptomyces ambofaciens ATCC 23877 [gbbct] | different | 0.853 | 0.252 | 0.292 | 0.299 |
| | | *INV domains* | | | | | |
| KJ825724 | *Heterodera glycines* | Heterodera glycines [gbinv] | similar | 1.052 | 0.802 | 0.762 | n/a |
| H_ave_c4372_g1_i2 | *Heterodera avenae* | Heterodera glycines [gbinv] | somewhat similar | 1.013 | 0.768 | 0.758 | n/a |
| G_ros_comp52539_c1_seq2 | *Globodera rostochiensis* | Heterodera glycines [gbinv] | similar | 1.032 | 0.796 | 0.771 | n/a |
| H_ave_c14716_g1_i1 | *Heterodera avenae* | Heterodera glycines [gbinv] | similar | 1.056 | 0.823 | 0.779 | n/a |
| G_ros_comp51573_c0_seq1 | *Globodera rostochiensis* | Heterodera glycines [gbinv] | similar | 1.014 | 0.795 | 0.784 | n/a |
| G_pal_J_comp3995_c1_seq1 | *Globodera pallida* | Heterodera glycines [gbinv] | somewhat similar | 1.000 | 0.783 | 0.783 | n/a |

## Table S3 continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| G_pal_7_comp88216_c0_seq2 | *Globodera pallida* | Heterodera glycines [gbinv] | similar | 1.026 | 0.792 | 0.772 | n/a |
| R_ren_c12967_g1_i1 | *Rotylenchus reneformis* | Heterodera glycines [gbinv] | similar | 1.016 | 0.813 | 0.8 | n/a |
| N_abe_c37491_g1_i1 | *Nacobbus aberrans* | Heterodera glycines [gbinv] | similar | 1.046 | 0.792 | 0.757 | n/a |
| N_abe_c37425_g1_i1 | *Nacobbus aberrans* | Heterodera glycines [gbinv] | different | 0.984 | 0.809 | 0.822 | n/a |
| N_abe_c37656_g1_i2 | *Nacobbus aberrans* | Heterodera glycines [gbinv] | similar | 1.029 | 0.829 | 0.806 | n/a |
| KJ825724 | *Heterodera glycines* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.956 | 0.511 | 0.532 | 0.537 |
| H_ave_c4372_g1_i2 | *Heterodera avenae* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.910 | 0.502 | 0.547 | 0.556 |
| G_ros_comp52539_c1_seq2 | *Globodera rostochiensis* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.934 | 0.508 | 0.539 | 0.549 |
| H_ave_c14716_g1_i1 | *Heterodera avenae* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.941 | 0.523 | 0.550 | 0.562 |
| G_ros_comp51573_c0_seq1 | *Globodera rostochiensis* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.940 | 0.554 | 0.594 | 0.585 |
| G_pal_J_comp3995_c1_seq1 | *Globodera pallida* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.930 | 0.549 | 0.589 | 0.592 |
| G_pal_7_comp88216_c0_seq2 | *Globodera pallida* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.938 | 0.555 | 0.592 | 0.591 |
| R_ren_c12967_g1_i1 | *Rotylenchus reneformis* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.929 | 0.602 | 0.647 | 0.649 |
| N_abe_c37491_g1_i1 | *Nacobbus aberrans* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.929 | 0.501 | 0.537 | 0.541 |
| N_abe_c37425_g1_i1 | *Nacobbus aberrans* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.914 | 0.647 | 0.708 | 0.707 |
| N_abe_c37656_g1_i2 | *Nacobbus aberrans* | Rhizobium leguminosarum bv. viciae 3841 [gbbct] | different | 0.924 | 0.646 | 0.702 | 0.696 |
| KJ825724 | *Heterodera glycines* | Rhizobium etli CFN 42 [gbbct] | different | 0.956 | 0.502 | 0.525 | 0.525 |
| H_ave_c4372_g1_i2 | *Heterodera avenae* | Rhizobium etli CFN 42 [gbbct] | different | 0.903 | 0.491 | 0.548 | 0.539 |
| G_ros_comp52539_c1_seq2 | *Globodera rostochiensis* | Rhizobium etli CFN 42 [gbbct] | different | 0.926 | 0.497 | 0.536 | 0.537 |
| H_ave_c14716_g1_i1 | *Heterodera avenae* | Rhizobium etli CFN 42 [gbbct] | different | 0.934 | 0.512 | 0.544 | 0.552 |
| G_ros_comp51573_c0_seq1 | *Globodera rostochiensis* | Rhizobium etli CFN 42 [gbbct] | different | 0.935 | 0.545 | 0.578 | 0.588 |
| G_pal_J_comp3995_c1_seq1 | *Globodera pallida* | Rhizobium etli CFN 42 [gbbct] | different | 0.932 | 0.540 | 0.579 | 0.580 |
| G_pal_7_comp88216_c0_seq2 | *Globodera pallida* | Rhizobium etli CFN 42 [gbbct] | different | 0.944 | 0.544 | 0.577 | 0.576 |
| R_ren_c12967_g1_i1 | *Rotylenchus reneformis* | Rhizobium etli CFN 42 [gbbct] | different | 0.925 | 0.591 | 0.643 | 0.635 |
| N_abe_c37491_g1_i1 | *Nacobbus aberrans* | Rhizobium etli CFN 42 [gbbct] | different | 0.924 | 0.491 | 0.532 | 0.531 |
| N_abe_c37425_g1_i1 | *Nacobbus aberrans* | Rhizobium etli CFN 42 [gbbct] | different | 0.915 | 0.637 | 0.695 | 0.698 |
| N_abe_c37656_g1_i2 | *Nacobbus aberrans* | Rhizobium etli CFN 42 [gbbct] | different | 0.925 | 0.638 | 0.694 | 0.686 |
| KJ825724 | *Heterodera glycines* | Sinorhizobium meliloti [gbbct] | different | 0.940 | 0.472 | 0.501 | 0.503 |
| H_ave_c4372_g1_i2 | *Heterodera avenae* | Sinorhizobium meliloti [gbbct] | different | 0.896 | 0.479 | 0.532 | 0.537 |
| G_ros_comp52539_c1_seq2 | *Globodera rostochiensis* | Sinorhizobium meliloti [gbbct] | different | 0.913 | 0.474 | 0.522 | 0.516 |
| H_ave_c14716_g1_i1 | *Heterodera avenae* | Sinorhizobium meliloti [gbbct] | different | 0.924 | 0.496 | 0.534 | 0.540 |
| G_ros_comp51573_c0_seq1 | *Globodera rostochiensis* | Sinorhizobium meliloti [gbbct] | different | 0.916 | 0.529 | 0.573 | 0.582 |
| G_pal_J_comp3995_c1_seq1 | *Globodera pallida* | Sinorhizobium meliloti [gbbct] | different | 0.912 | 0.524 | 0.576 | 0.573 |
| G_pal_7_comp88216_c0_seq2 | *Globodera pallida* | Sinorhizobium meliloti [gbbct] | different | 0.928 | 0.541 | 0.583 | 0.583 |
| R_ren_c12967_g1_i1 | *Rotylenchus reneformis* | Sinorhizobium meliloti [gbbct] | different | 0.918 | 0.585 | 0.635 | 0.640 |
| N_abe_c37491_g1_i1 | *Nacobbus aberrans* | Sinorhizobium meliloti [gbbct] | different | 0.911 | 0.475 | 0.525 | 0.518 |
| N_abe_c37425_g1_i1 | *Nacobbus aberrans* | Sinorhizobium meliloti [gbbct] | different | 0.910 | 0.645 | 0.713 | 0.705 |
| N_abe_c37656_g1_i2 | *Nacobbus aberrans* | Sinorhizobium meliloti [gbbct] | different | 0.912 | 0.644 | 0.704 | 0.709 |
| | | *CM domains* | | | | | |
| KJ825727 | *Heterodera glycines* | Heterodera glycines [gbinv] | similar | 1.057 | 0.854 | 0.808 | n/a |

## Table S3 continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| AAO19577.2 | *Heterodera glycines* | Heterodera glycines [gbinv] | similar | 1.034 | 0.818 | 0.791 | n/a |
| ABA06538.1 | *Heterodera schachtii* | Heterodera glycines [gbinv] | similar | 1.047 | 0.845 | 0.807 | n/a |
| g2131.t1_S_28_200000 | *Globodera rostochiensis* | Heterodera glycines [gbinv] | similar | 1.045 | 0.790 | 0.756 | n/a |
| ABR19887.1 | *Globodera rostochiensis* | Heterodera glycines [gbinv] | somewhat similar | 1.007 | 0.818 | 0.812 | n/a |
| G_pal_7_comp26511_c0_seq2 | *Globodera pallida* | Heterodera glycines [gbinv] | similar | 1.020 | 0.773 | 0.758 | n/a |
| AEA07500.1 | *Globodera pallida* | Heterodera glycines [gbinv] | similar | 1.047 | 0.846 | 0.808 | n/a |
| AEA07501.1 | *Globodera tabacum* | Heterodera glycines [gbinv] | similar | 1.018 | 0.829 | 0.814 | n/a |
| AIE45298.1 | *Globodera ellingtonae* | Heterodera glycines [gbinv] | similar | 1.037 | 0.833 | 0.803 | n/a |
| R_ren_comp46249_c2_seq2 | *Rotylenchus reneformis* | Heterodera glycines [gbinv] | different | 0.975 | 0.780 | 0.800 | n/a |
| R_ren_comp43183_c0_seq3 | *Rotylenchus reneformis* | Heterodera glycines [gbinv] | similar | 1.020 | 0.851 | 0.834 | n/a |
| N_abe_c34149_g1_i1 | *Nacobbus aberrans* | Heterodera glycines [gbinv] | somewhat similar | 0.996 | 0.814 | 0.817 | n/a |
| KJ825727 | *Heterodera glycines* | Burkholderia pseudomallei 668 [gbbct] | different | 0.918 | 0.439 | 0.476 | 0.48 |
| AAO19577.2 | *Heterodera glycines* | Burkholderia pseudomallei 668 [gbbct] | different | 0.954 | 0.405 | 0.428 | 0.421 |
| ABA06538.1 | *Heterodera schachtii* | Burkholderia pseudomallei 668 [gbbct] | different | 0.933 | 0.415 | 0.446 | 0.444 |
| g2131.t1_S_28_200000 | *Globodera rostochiensis* | Burkholderia pseudomallei 668 [gbbct] | different | 0.948 | 0.313 | 0.328 | 0.332 |
| ABR19887.1 | *Globodera rostochiensis* | Burkholderia pseudomallei 668 [gbbct] | different | 0.856 | 0.456 | 0.541 | 0.524 |
| G_pal_7_comp26511_c0_seq2 | *Globodera pallida* | Burkholderia pseudomallei 668 [gbbct] | different | 0.931 | 0.297 | 0.318 | 0.320 |
| AEA07500.1 | *Globodera pallida* | Burkholderia pseudomallei 668 [gbbct] | different | 0.879 | 0.443 | 0.507 | 0.501 |
| AEA07501.1 | *Globodera tabacum* | Burkholderia pseudomallei 668 [gbbct] | different | 0.867 | 0.472 | 0.542 | 0.547 |
| AIE45298.1 | *Globodera ellingtonae* | Burkholderia pseudomallei 668 [gbbct] | different | 0.911 | 0.457 | 0.502 | 0.501 |
| R_ren_comp46249_c2_seq2 | *Rotylenchus reneformis* | Burkholderia pseudomallei 668 [gbbct] | different | 0.838 | 0.363 | 0.431 | 0.435 |
| R_ren_comp43183_c0_seq3 | *Rotylenchus reneformis* | Burkholderia pseudomallei 668 [gbbct] | different | 0.861 | 0.478 | 0.552 | 0.558 |
| N_abe_c34149_g1_i1 | *Nacobbus aberrans* | Burkholderia pseudomallei 668 [gbbct] | different | 0.839 | 0.482 | 0.572 | 0.577 |
| KJ825727 | *Heterodera glycines* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.924 | 0.454 | 0.495 | 0.488 |
| AAO19577.2 | *Heterodera glycines* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.949 | 0.412 | 0.435 | 0.433 |
| ABA06538.1 | *Heterodera schachtii* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.933 | 0.423 | 0.455 | 0.452 |
| g2131.t1_S_28_200000 | *Globodera rostochiensis* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.945 | 0.318 | 0.338 | 0.335 |
| ABR19887.1 | *Globodera rostochiensis* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.862 | 0.465 | 0.541 | 0.538 |
| G_pal_7_comp26511_c0_seq2 | *Globodera pallida* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.923 | 0.301 | 0.330 | 0.322 |
| AEA07500.1 | *Globodera pallida* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.900 | 0.454 | 0.503 | 0.506 |
| AEA07501.1 | *Globodera tabacum* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.884 | 0.482 | 0.548 | 0.542 |
| AIE45298.1 | *Globodera ellingtonae* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.915 | 0.466 | 0.510 | 0.509 |
| R_ren_comp46249_c2_seq2 | *Rotylenchus reneformis* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.863 | 0.379 | 0.443 | 0.435 |
| R_ren_comp43183_c0_seq3 | *Rotylenchus reneformis* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.895 | 0.501 | 0.559 | 0.560 |
| N_abe_c34149_g1_i1 | *Nacobbus aberrans* | Burkholderia cenocepacia HI2424 [gbbct] | different | 0.855 | 0.497 | 0.582 | 0.581 |
| KJ825727 | *Heterodera glycines* | Burkholderia thailandensis E264 [gbbct] | different | 0.922 | 0.440 | 0.471 | 0.483 |
| AAO19577.2 | *Heterodera glycines* | Burkholderia thailandensis E264 [gbbct] | different | 0.937 | 0.404 | 0.426 | 0.436 |
| ABA06538.1 | *Heterodera schachtii* | Burkholderia thailandensis E264 [gbbct] | different | 0.931 | 0.413 | 0.444 | 0.443 |
| g2131.t1_S_28_200000 | *Globodera rostochiensis* | Burkholderia thailandensis E264 [gbbct] | different | 0.953 | 0.314 | 0.327 | 0.332 |
| ABR19887.1 | *Globodera rostochiensis* | Burkholderia thailandensis E264 [gbbct] | different | 0.858 | 0.457 | 0.536 | 0.529 |

Table S3 continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| G_pal_7_comp26511_c0_seq2 | *Globodera pallida* | Burkholderia thailandensis E264 [gbbct] | different | 0.933 | 0.299 | 0.319 | 0.322 |
| AEA07500.1 | *Globodera pallida* | Burkholderia thailandensis E264 [gbbct] | different | 0.887 | 0.444 | 0.501 | 0.500 |
| AEA07501.1 | *Globodera tabacum* | Burkholderia thailandensis E264 [gbbct] | different | 0.874 | 0.473 | 0.540 | 0.542 |
| AIE45298.1 | *Globodera ellingtonae* | Burkholderia thailandensis E264 [gbbct] | different | 0.916 | 0.458 | 0.500 | 0.500 |
| R_ren_comp46249_c2_seq2 | *Rotylenchus reneformis* | Burkholderia thailandensis E264 [gbbct] | different | 0.845 | 0.365 | 0.432 | 0.432 |
| R_ren_comp43183_c0_seq3 | *Rotylenchus reneformis* | Burkholderia thailandensis E264 [gbbct] | different | 0.860 | 0.478 | 0.557 | 0.554 |
| N_abe_c34149_g1_i1 | *Nacobbus aberrans* | Burkholderia thailandensis E264 [gbbct] | different | 0.839 | 0.484 | 0.579 | 0.575 |

The listed codon usage tables are available at http://www.kazusa.or.jp/codon/. Codon Adaptation Index (CAI) was calcuated using the CAIcal server (Puigbo et al. 2008). CAI ranges from 0-1 with 0 corresponding to no similarity in codon usage to the respective table and 1 corresponding to identical codon usage. Expected CAI (E-CAI) was calculated using the E-CAI server (Puigbo et al. 2008). E-CAI corresponds to the upper one-tailed limit of a 95% confidence interval containing 95% of the population of CAI values calculated from 500 simulated sequences containing the same amino acid composition and %GC content as the input sequence. The Markov method was used to estimate the E-CAI value. For Hoplolaimina compared to bacterial tables both standard and eubacterial genetic codes were used. When both standard and eubacterial genetic codes were used, the average E-CAI was used for CAI/E-CAI calculations. In general, CAI/E-CAI values greater than 1 indicate that the input sequence contains significantly similar codon usage with the compared codon usage table. However, we found that the E-CAI estimate can vary up to 0.010, and thus when CAI and E-CAI values were within 0.010, we concluded this as 'somewhat similar'. But when CAI was greater than E-CAI by more than 0.010 we concluded this as significantly 'similar'. When CAI was less than E-CAI by more than 0.010 we concluded this as 'different' codon usage. We chose to implement only the Heterodera glycines [gbinv] codon usage table as it contains far more codons than any other Hoplolaimina table availalbe at http://www.kazusa.or.jp/codon/. Furthermore, since we were unable to implement a codon usage table for *Meloidogyne incognita*, we did not evaluate the *M. incognita* sequences. *Nacobbus aberrans* was evaluated only because its %GC content was found to be comparable to Heteroderidae and Hoplolaimidae (Fig. 7). Finally, to avoid possible bias towards a single bacterial species, we implemented three species per donor bacterium.

## Supplementary Text

### Evolution of *GNATs*, *INVs* and *CMs* in Hoplolaimina following HGT from rhizosphere bacteria

The family of *GLAND1*s (*FAM7 GNATs*) appeared to be rather simple in that, as mentioned in the main text, it seemed to have originated via horizontal gene transfer (HGT) in Hoplolaimina clade 2B (Fig. 2A). We only identified a single homolog in each Hoplolaimina plant-parasitic nematode (PPN; cyst nematodes and reniform nematode *Rotylenchulus reniformis*) (see Fig. S2), and with their highly significant sequence similarities (Fig. S6), this strongly suggested orthology. Also, the protein sequences were found to be highly similar within and near their GNAT domains, and signal peptides were identified at the N-termini of all orthologs (Fig. S6), suggesting that all may be effectors

in the respective Hoplolaimina PPN. On the other hand, sequences N- and C-terminal to the GNAT domain in all orthologs are highly variable, suggesting that these regions of the protein are evolving relatively fast. Given that *Hg-GLAND1* was found to encode a candidate effector (Noon et al. 2015), the latter finding led us to speculate that those regions might be involved in binding to target proteins, possibly for acetylation.

We next inspected the *invertase* (*INV*) gene family. In the raw phylogenetic tree shown in Figure S3, from which we constructed a subtree shown in Figure S7, a complexity of gene duplication and loss was evident (however we consider that gene loss may simply be an artifact of incomplete genome assemblies or lack of detected expression from the evaluated transcriptomes). For instance, three different, highly supported clusters were found (Fig. S7). Cluster 1 is in a nested, highly supported monophyletic group with the donor bacteria Rhizobiales and consists of orthologs—indicated by bootstrap values > 95—from *Heterodera glycines*, *Heterodera avenae*, *Globodera pallida* and *Nacobbus aberrans*, strongly suggesting that this cluster contains the original transferred genes (i.e., xenologs). Interestingly, *G. pallida* and *N. aberrans* INVs in cluster 1 are the homologs that resulted in %GC contents significantly greater than expected and within the range of Rhizobiales (Fig. 7B), further supporting xenology of cluster 1 with Rhizobiales INVs.

Clusters 2 and 3 are in a monophyletic group, and although not highly supported (bootstrap of only 59), suggest two duplications (Fig. S7). The first duplication resulted in clusters 1 and 2/3 and probably occurred prior to divergence of Hoplolaimina clade 2 due to the presence of both Hoplolaimina clade 2A and 2B PPN in these clusters (see Fig. 1B). The second duplication probably occurred in the cyst nematode lineage since only

*H. glycines* and *G. rostochiensis* are found in cluster 3, which contains the *GLAND13s*. Also, we identified another duplication within *M. incognita* (cluster 2b; Fig. S7), which probably occurred within the root-knot nematode lineage and before its divergence due to the relatively long branch lengths (sequence divergence)—additional root-knot nematode sequences could be analyzed for that duplication in future studies. These interpretations would likewise propose the following gene losses: First, the cluster 1 orthologs, and thus the likely xenologs of the rhizobacterial *INVs*, may have been lost in root-knot nematodes (at least in *M. incognita*), *R. reniformis* and *G. rostochiensis*. Second, the cluster 3 *GLAND13* may have been lost from *G. pallida* and *H. avenae*, and possibly other *Globodera* and *Heterodera* species, with no apparent losses from cluster 2 (Fig. S7). As mentioned above, it is still possible that some of these alleged gene losses are actually due to insufficient sequence data.

As shown in Figure S7, all INV homologs were found to consist of predominantly a single GH32 functional domain (InterPro: IPR001362). On the other hand, we were surprised to find extensive variation from signal peptide predictions (see Materials and Methods). The following three classes were found: First, we found probable secreted forms containing predicted N-terminal signal peptides with no overlap with transmembrane (TM) regions, which contained both GLAND13s, consistent with their candidacy as effectors, as well as the Ha-INV from cluster 1 (Fig. S7). Second, homologs were found to be uncertain whether secreted or TM forms due to overlapping predictions of signal peptides and TM regions, which contained over half of the GH32 homologs (Fig. S7). Third, and of particular interest, we found probable TM forms containing predicted, non-N-terminal TM regions without predicted signal peptides. Firstly, class 2

is not an uncommon challenge for signal peptide predictions as both signal peptides and TM regions contain similar stretches of hydrophobic amino acids; referred to as the H-region in signal peptides (Kall et al. 2007). We found that these class 2 INVs are all positive for predicted signal peptides only when the assumption is provided that TM regions are absent, but not when TM regions are considered, and all are predicted positive for TM regions regardless of criteria, further adding to the uncertainty. However, the TM regions predicted in all class 2 homologs overlap almost perfectly where the signal peptide H-region is typically found (Fig. S7). Thus, we believed that it was more likely that these are degenerate signal peptides, and therefore also possible secreted forms, rather than TM regions, but this of course will have to be empirically tested in future studies to be certain.

The particular interest in the class 3 INVs (Fig. S7) is threefold: Firstly, these probable TM forms are only found in false root-knot nematodes and root-knot nematodes, and thus appear to have formed in Hoplolaimina clade 2A (see Fig. 1B). Secondly, the class 3 Mi-INV contains a GH32 domain with an amino acid/polyamine transporter 1 (aka AA permease) domain (InterPro: IPR002293) appended at the N-terminus. The AA permease domain comprises almost the entire N-terminal half of the protein and contains at least nine TM regions. Also, in a separate blastp search with the Mi-INV AA permease domain as query, we identified highly significant matches to numerous AA permeases from nematodes, of which all were found to consist entirely of single AA permease domains (data not shown). Noteworthy, this class 3 Mi-INV (Minc02870) is the highly diverged *M. incognita* paralog in cluster 2b (Fig. S7). Taken together, this suggested that in the root-knot nematode lineage, the GH32 domain

acquired from rhizobacteria was appended to an AA permease domain from the nematode genome. Thirdly, although the GH32 domains from the class 3 *N. aberrans* homologs are predicted as non-cytoplasmic suggesting extracellular processing of sucrose within the nematode, the GH32 domain from the *M. incognita* homolog is predicted to be on the cytoplasmic surface (data not shown). Accordingly, we speculate that the class 3 Mi-INV might use a cotransport mechanism (either symport or antiport) characteristic of AA permeases. Such a cotransport mechanism would facilitate transport of sucrose into the cytoplasm of the respective cells, where the GH32 functional domain may then process sucrose into glucose and fructose to provide nourishment for the nematode. Empirical testing of the latter hypothesis will be of particular interest for future studies of root-knot nematode INVs.

Lastly, for the Hoplolaimina *chorismate mutase* (*CM*) gene family, a subtree was generated from the raw phylogenetic tree shown in Figure S4, and is presented in Figure S8. We observed four clusters of Hoplolaimina *CMs*. Clusters 1 (little support; boostrap of only 57) and 2 were composed of the root-knot nematode *Meloidogyne arenaria* and the false root-knot nematode *N. aberrans*, and exclusively root-knot nematodes, respectively, and thus Hoplolaimina clade 2A (see Fig. 1B). Clusters 3 and 4 contained cyst nematodes and *Ro. reniformis*, and thus Hoplolaimina clade 2B (Fig. S8; see Fig. 1B for reference). The overall pattern of the clusters suggested that clusters 1, 2 and 3 contain the xenologs of the *Burkholderia*-related *CMs*, while there appeared to be a duplication of *CMs* in Hoplolaimina clade 2B that resulted in the *GLAND16s* in cluster 4. Since there is little support for cluster 1, and cluster 1's grouping with *Burkholderia* was not at all supported, in addition to the finding that the grouping of clusters 2 and 3/4 is

not supported, we could not determine which *M. arenaria* paralog (revealing another duplication) is the more likely xenolog.

As mentioned throughout the main text, the relative sizes of GLAND16s are interesting (Fig. S8, we could not obtain the complete sequences for the *H. avenae* and *Ro. reniformis* orthologs, indicated by dots in the C-termini), and as mentioned above, the gene forms first appeared in the Hoplolaimina clade 2B lineage (see Fig. 1B). Interestingly, although in a blastp search against the non-redundant (NR) database using the amino acids that flanked the GLAND16 CM domains we did not identify any similar sequences, in a separate search of all nematode genome and transcriptome sequences at Nematode.net we identified significant matches to novel sequences from numerous nematodes (see Materials and Methods, and data not shown). This latter finding suggested that the *CM* domain derived from *Burkholderia*-related bacteria was inserted into a nematode gene resulting in *GLAND16s* (Fig. S8), somewhat similar to what we observed for the class 3 *Mi-INV* (Fig. S7). Another interesting finding, although predicted signal peptides were found at the N-termini of nearly all CM homologs, this was not observed for *N. aberrans* (Fig. S8). Na-CM does not contain a signal peptide, but rather, two TM regions in the N-terminal half of the protein. Given that the shikimate pathway is believed to be absent from animals, and thus nematodes, but present in plants, Na-CM not being secreted into the plant, but rather, likely functioning within the nematode is somewhat paradoxical. It will be interesting to see in future studies what the significances are for the evident subfunctionalizations of both the INVs and CMs in particular Hoplolaimina PPN.

# APPENDIX C. CHAPTER 4 SUPPLEMENTARY INFORMATION

## Supplementary Tables

Note that Table S1 will be available online at the journal's website upon acceptance for publication. This table presents raw data that was determined to be inessential to the presentation of results for this dissertation. Table S2 presents data pertinent to the presentation of results for this dissertation and is provided below.

**Table S2. Nematode genomes tblastn-searched for Phlebovirus glycoprotein G2**

| Nematode species | Resource | Accession/ID | Reference |
|---|---|---|---|
| *Heterodera glycines* | in house | . | Hudson and Lambert, unpublished |
| *Globodera pallida* | Sanger web portal | . | Cotton et al. (2014) |
| *Meloidogyne incognita* | NCBI | GCA_000180415.1 | Abad et al. (2008) |
| *Meloidogyne incognita* | INRA web portal | CNS 2007-10 | . |
| *Meloidogyne hapla* | NCBI | GCA_000172435.1 | Opperman et al. (2008) |
| *Bursaphelenchus xylophilus* | NCBI | GCA_000231135.1 | Kikuchi et al. (2011) |
| *Ascaris suum* | NCBI | GCA_000298755.1 | Jex et al. (2011) |
| *Brugia malayi* | NCBI | GCA_000002995.1 | Ghedin and Anton (2009) |
| *Caenorhabditis briggsae* | NCBI | GCA_000004555.3 | . |
| *Caenorhabditis elegans* | Ensemble web portal | . | Consortium (1998) |
| *Caenorhabditis remanei* | NCBI | GCA_000149515.1 | . |
| *Haemonchus contortus* | NCBI | GCA_000469685.1 | Schwarz et al. (2013) |
| *Heterorhabditis bacteriophora* | NCBI | GCA_000223415.1 | Bai et al. (2013) |
| *Loa loa* | NCBI | GCF_000183805.1 | Desjardins et al. (2013); Tallon et al. (2014) |
| *Necator americanus* | NCBI | GCA_000507365.1 | Tang et al. (2014) |
| *Pristionchus pacificus* | NCBI | GCA_000180635.1 | Dieterich et al. (2008) |
| *Trichinella spiralis* | NCBI | GCA_000181795.2 | Mitreva et al. (2011) |
| *Trichuris suis* | NCBI | GCA_000701005.1 | Jex et al. (2014) |
| *Trichuris trichiura* | NCBI | GCA_000613005.1 | Foth et al. (2014) |

# APPENDIX D. CHAPTER 5 SUPPLEMENTARY INFORMATION

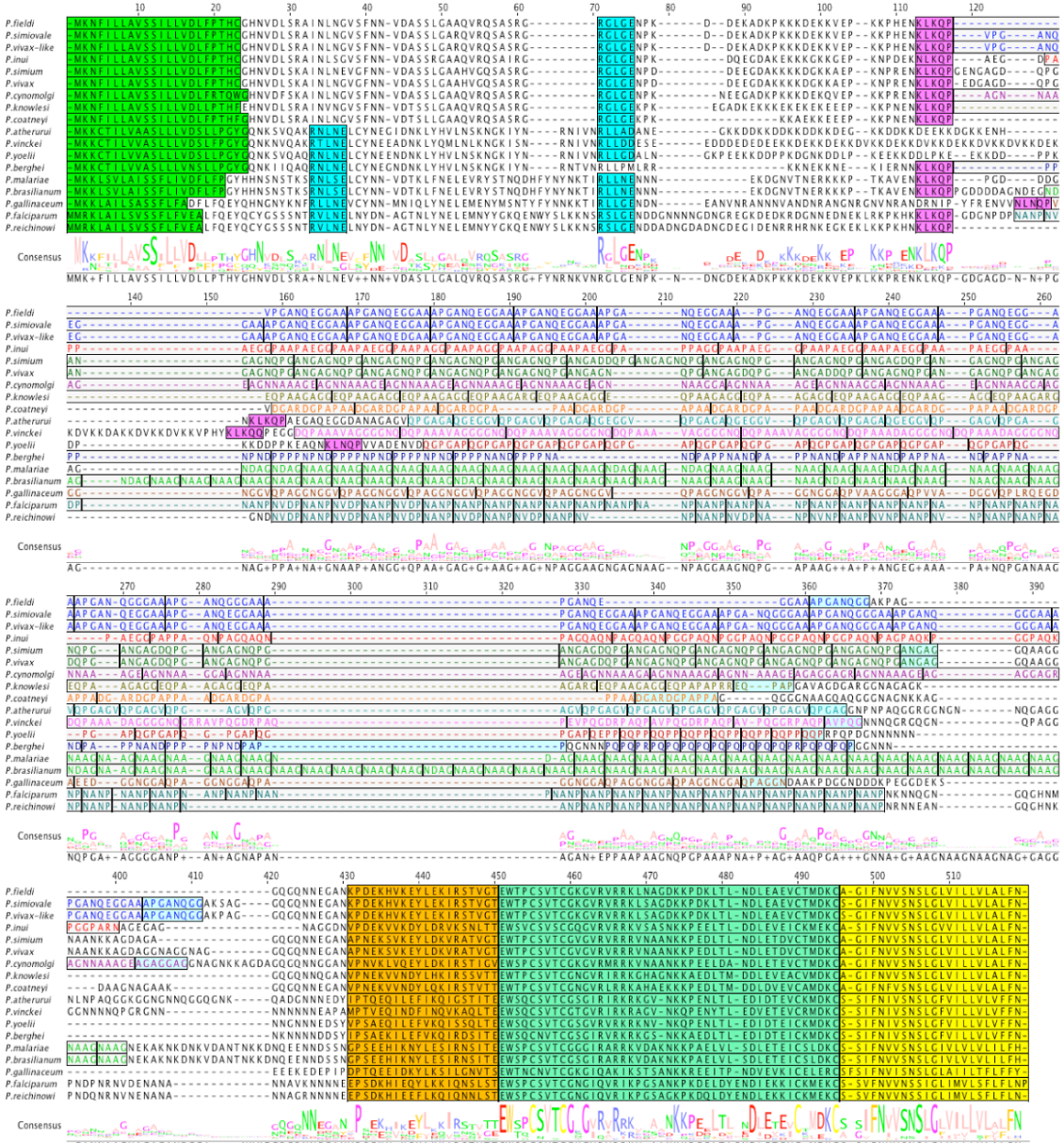


**Fig. S1. Multiple sequence alignment of *Plasmodium* CSPs and illustration of domains.** The multiple sequence alignment (MSA) was generated for the eighteen *Plasmodium* CSPs whose protein sequences are available in GenBank using the MUSCLE algorithm and was evaluated, and annotated, with Jalview. Signal peptides are highlighted green. *Plasmodium* export element/vacuolar targeting signal (PEXEL/VTS) motifs are highlighted cyan. The five most conserved amino acids of the region I (RI) domain are highlighted purple, but importantly, approximately nine upstream amino acids are also part of the larger RI domain. The species-specific repetitive region (RR) is

highlighted light gray with the amino acids within the RR domains that are identical between some closely related species colored accordingly. Region III (RIII) is highlighted orange, region II+ (RII+) is highlighted aquamarine, and the glycosylphosphatidylinositol (GPI)-anchor is highlighted yellow. A consensus sequence is provided below the MSA.

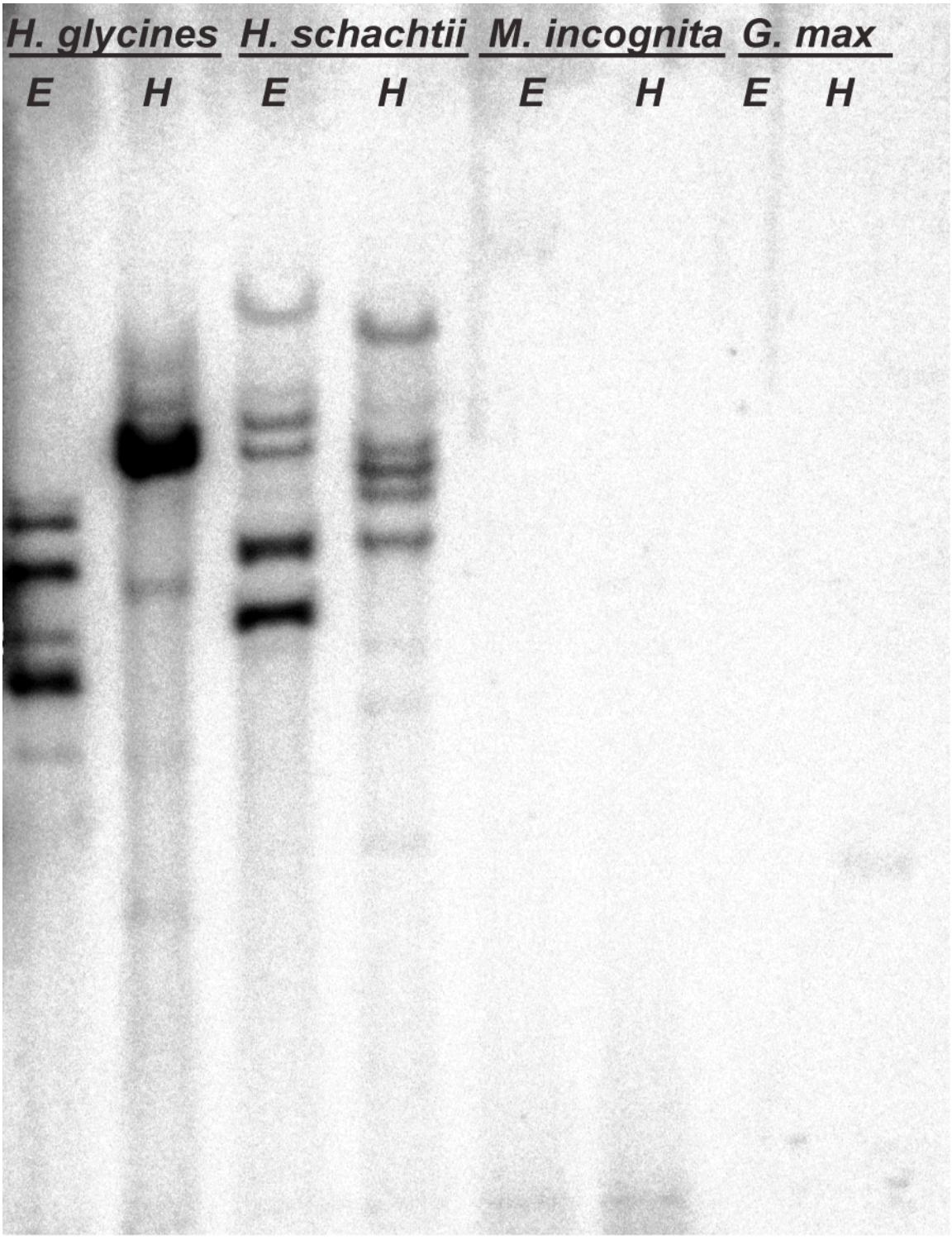



**Fig. S2. *GLAND18* Southern blot.** Genomic (g)DNA was isolated from *Heterodera glycines*, *Heterodera schachtii*, *Meloidogyne incognita* and soybean, digested with EcoRI (E) and HindIII (H) restriction endonucleases and 10-µg was electrophoresed on a 2% agarose gel and subjected to Southern analysis using an HgGLAND18 CDS $^{32}$P-labeled DNA probe. The DNA gel blot was imaged via phosphorimaging. *H. glycines*, and *M.*

*incognita* and soybean gDNA, were included as positive and negative controls, respectively.

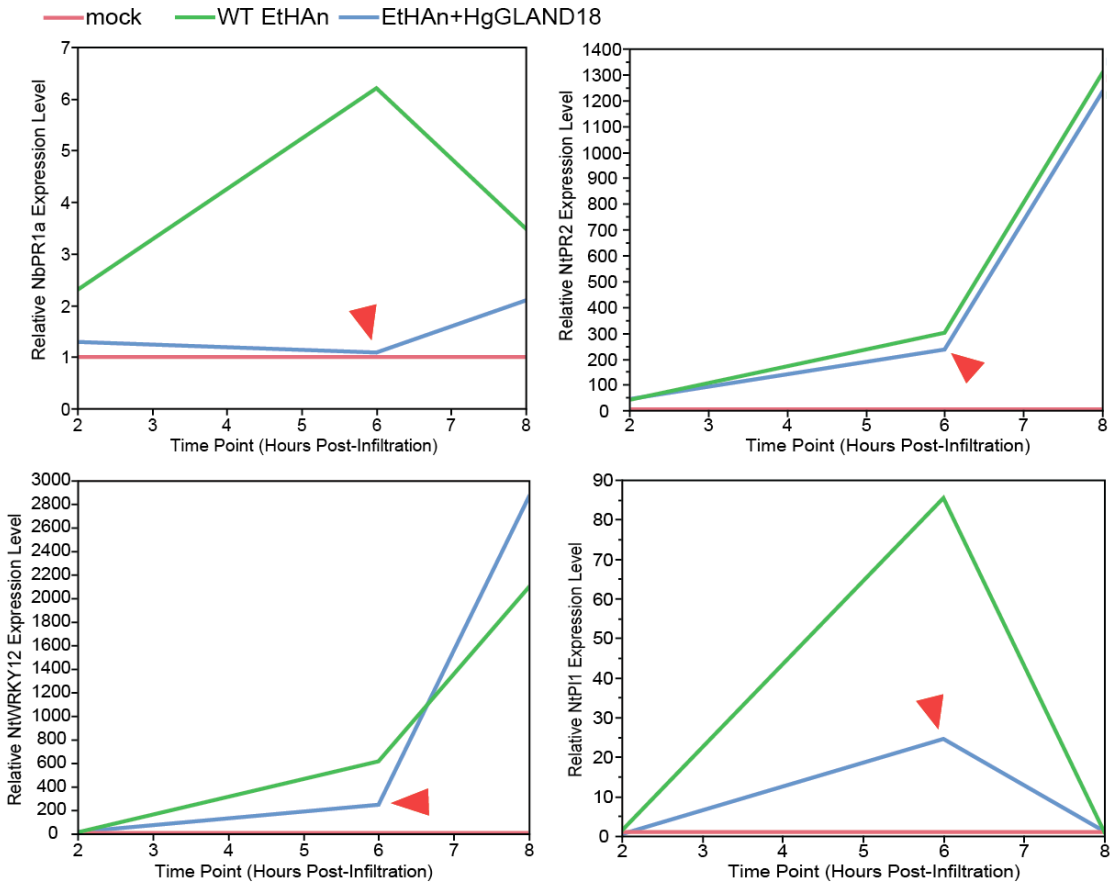



**Fig. S3. qRT-PCR screen for the optimum time point for quantification of salicylic acid-responsive defense marker gene expression during PTI responses.** qRT-PCR was performed on *Nicotiana benthamiana* leaves infiltrated with either wild type (WT) *Pseudomonas fluorescens* strain EtHAn, EtHAn + HgGLAND18$^{-sp}$, or 10-mM $MgCl_2$ buffer (mock) in order to quantify changes in relative expression levels for salicylic acid (SA)-responsive defense marker genes. Express levels are relative to mock. Data were normalized to *NbActin1*. Time points evaluated were 2-, 4-, 6- and 8-hrs post-infiltration (hpi), and a single biological replicate was used for each in the screen. The optimum time point was 6-hpi, which resulted in the maximum difference between WT EtHAn and EtHAn + HgGLAND18$^{-sp}$ (red arrow head), and was used in Figure 4E.

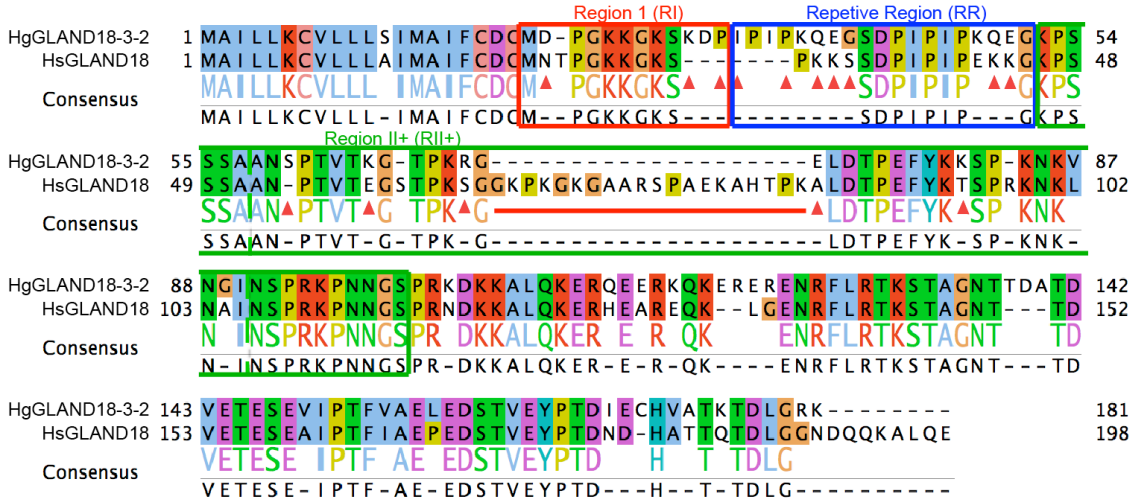



**Fig. S4. Pairwise sequence alignment of GLAND18 protein sequences from *Heterodera glycines* and *Heterodera schachtii*.** The pairwise sequence alignment of HgGLAND18 (isoform 3-2) and HsGLAND18 was generated with the MUSCLE algorithm using default parameters in the MEGA6 program, and was evaluated with Jalview. A consensus sequence is provided below the alignment. Amino acids within HgGLAND18 that are similar to *Plasmodium* RI, RR and RII+ CSP domains are annotated accordingly. Insertions/deletions within HsG18 that disrupted the similarity that was observed for HgGLAND18 to *Plasmodium* RI, RR and RII+ CSP domains are shown with red arrowheads, and a large insertion within HsGLAND18 that largely disrupted the similarity to RII+ is underlined red.

**Table S1. GenBank accession numbers for all *Plasmodium* CSP sequences mentioned in our study**

| ***Plasmodium* Species** | **Accession Number** |
|---|---|
| *P. falciparum* | AAA63422.1 |
| *P. knowlesi* | AAA29540.1 |
| *P. yoelii* | AAA29558.1 |
| *P. inui* | CBI14799.1 |
| *P. malariae* | CAA05624.1 |
| *P. gallinaceum* | AAC47344.1 |
| *P. reichinowi* | AAA29561.1 |
| *P. cynomolgi* | BAM14085.1 |
| *P. fieldi* | AFD97256.1 |
| *P. simiovale* | AAA18617.1 |
| *P. vivax*-like | AAA18616.1 |
| *P. atheruri* | AAK56911.1 |
| *P. coatneyi* | ADN94531.1 |
| *P. berghei* | AAA29577.1 |
| *P. vinckei* | AAL36457.1 |
| *P. simium* | AAA70381.1 |
| *P. vivax* | ADB92538.1 |
| *P. brasilianum* | P14593.1 |

**Table S2. Model selection analyses for HgGLAND18 and controls with *Plasmodium* CSPs**

| Evolutionary Model | Substitution Model | No. of Parameters | ΔBIC | ΔAICc |
|---|---|---|---|---|
| *Test 1, HgGLAND18 vs. Bacillus proteins and Plasmodium CSPs* | | | | |
| HgGLAND18+Bacillus proteins, Plasmodium CSPs | GTR+G | 52 | 0 | 0 |
| HgGLAND18+Plasmodium CSPs, Bacillus proteins | GTR+G | 52 | 19.913 | 19.914 |
| *Test 2, HgGLAND18 and controls vs. Plasmodium* | | | | |
| HgGLAND18, Plasmodium CSPs | GTR+G+I | 45 | 0 | 0 |
| HgGLAND18+Plasmodium CSPs | GTR+G | 44 | 20.068 | 14.307 |
| HgGLAND18+Asian Primate malaria CSPs, falciparum/reichinowi+Avian+Reptile+Rodent+African Primate malaria CSPs | GTR+G | 44 | 29.763 | 24.002 |
| HgGLAND18+P. fieldi/simiovale/vivax-like, falciparum/reichinowi+Avian+Reptile+Rodent+Primate malaria CSPs | GTR+G | 44 | 43.402 | 37.542 |
| HgGLAND8, Plasmodium CSPs | GTR+G+I | 45 | 0 | 0 |
| HgGLAND8+Plasmodium CSPs | GTR+G+I | 45 | -0.173 | -0.173 |
| HgGLAND8+Asian Primate malaria CSPs, falciparum/reichinowi+Avian+Reptile+Rodent+African Primate malaria CSPs | GTR+G+I | 45 | -0.554 | -0.555 |
| HgGLAND8+P. fieldi/simiovale/vivax-like, falciparum/reichinowi+Avian+Reptile+Rodent+Primate malaria CSPs | GTR+G | 44 | 6.778 | 0.846 |
| B. cereus 'CSP', Plasmodium CSPs | GTR+G | 44 | 0 | 0 |
| B. cereus 'CSP'+Plasmodium CSPs | GTR+G | 44 | -0.232 | -0.231 |
| B. cereus 'CSP'+Asian Primate malaria CSPs, falciparum/reichinowi+Avian+Reptile+Rodent+African Primate malaria CSPs | GTR+G | 44 | -1.831 | -1.831 |
| B. cereus 'CSP'+P. fieldi/simiovale/vivax-like, falciparum/reichinowi+Avian+Reptile+Rodent+Primate malaria CSPs | GTR+G | 44 | -2.273 | -2.272 |
| Human SARMP2, Plasmodium CSPs | GTR+G | 44 | 0 | 0 |
| Human SARMP2+Plasmodium CSPs | GTR+G | 44 | 0.24 | 0.239 |
| Human SARMP2+Asian Primate malaria CSPs, falciparum/reichinowi+Avian+Reptile+Rodent+African Primate malaria CSPs | GTR+G | 44 | 0.293 | 0.292 |
| Human SARMP2+P. fieldi/simiovale/vivax-like, falciparum/reichinowi+Avian+Reptile+Rodent+Primate malaria CSPs | GTR+G | 44 | 0.043 | 0.042 |
| P. falciparum EMP1, Plasmodium CSPs | GTR+G | 44 | 0 | 0 |
| P. falciparum EMP1+Plasmodium CSPs | GTR+G | 44 | 0.734 | 0.734 |
| P. falciparum EMP1+Asian Primate malaria CSPs, falciparum/reichinowi+Avian+Reptile+Rodent+African Primate malaria CSPs | GTR+G | 44 | 1.271 | 1.27 |
| P. falciparum EMP1+P. fieldi/simiovale/vivax-like, falciparum/reichinowi+Avian+Reptile+Rodent+Primate malaria CSPs | GTR+G | 44 | 1.286 | 0.978 |
| HsGLAND18, Plasmodium CSPs | GTR+G | 44 | 0 | 0 |
| HsGLAND18+Plasmodium CSPs | GTR+G | 44 | -0.211 | -0.212 |
| HsGLAND18+Asian Primate malaria CSPs, falciparum/reichinowi+Avian+Reptile+Rodent+African Primate malaria CSPs | GTR+G | 44 | -1.44 | -1.44 |
| HsGLAND18+P. fieldi/simiovale/vivax-like, falciparum/reichinowi+Avian+Reptile+Rodent+Primate malaria CSPs | GTR+G | 44 | -2.13 | -2.15 |

The first analysis (test 1) tested whether the HgGLAND18 N-terminal (CSP-like) domain was more likely related to *Bacillus* proteins (HgGLAND18+*Bacillus* proteins, *Plasmodium* CSPs) or *Plasmodium* CSPs (HgGLAND18+*Plasmodium* CSPs, *Bacillus* proteins). The second analysis tested whether the HgGLAND18 N-terminal (CSP-like)

domain was more likely to be unrelated to *Plasmodium* CSPs (HgGLAND18, *Plasmodium* CSPs) or related (HgGLAND18+*Plasmodium* CSPs), with additional testing for specific clades and monophyletic groups within *Plasmodium*. The second analysis also tested the following five different controls (Materials and Methods): HgGLAND8, *Bacillus cereus* 'circumsporozoite protein' ('CSP'), Human SARMP2, *Plasmodium falciparum* EMP1, and *Heterodera schachtii* GLAND18 (HsGLAND18). Also included in the table are the best scoring models of amino acids substitution and the number of parameters used, and the resulting Bayesian and corrected Akaike Information Criteria (BIC and AICc) scores expressed as the difference (Δ) from the topmost model in each analysis. Score differences of greater than 5 are considered strong empirical evidence for the better scoring evolutionary model (Materials and Methods). In the table, higher scores are better. Evolutionary models that resulted in differences greater than 5 are highlighted yellow.

**Table S3. Plant-parasitic nematode raw sequence reads searched for HgGLAND18 homologs**

| Species | Sequence reads | Reference |
|---|---|---|
| *Heterodera avenae* | transcriptomic | Kumar et al. (2014) |
| *Globodera pallida* | transcriptomic and genomic | Cotton et al. (2014) |
| *Globodera rostochiensis* | transcriptomic and genomic | Eves-van den Akker et al., unpublished |
| *Globodera ellingtonae* | transcriptomic and genomic | Phillips et al., unpublished |
| *Nacobbus aberrans* | transcriptomic | Eves-van den Akker et al. (2014) |
| *Rotylenchulus reneformis* | transcriptomic | Eves-van den Akker et al., unpublished |
| *Hirschmaniella oryzae* | transcriptomic | Eves-van den Akker et al., unpublished |
| *Longidorus elongatus* | transcriptomic | Jones et al., unpublished |

## Table S4. Complete list of primers used in our study

| Primer Name | Primer Sequence (5'-3') | Application |
|---|---|---|
| HgG18_repeats_F | TCATGGCGATTTTCTGCGACTGT | RT-PCR on *H. glycines* |
| HgG18_repeats_R | CGTTTCGGAGTGCCTTTTGTTACTGTC | RT-PCR on *H. glycines* |
| HgActin_RT-PCR_ref_F | AAGGCCAACAGAGAAAAGATGAC | RT-PCR on *H. glycines* |
| HgActin_RT-PCR_ref_R | TTCATCAGGTAGTCAGTGAGGTC | RT-PCR on *H. glycines* |
| HgG18_RT-PCR_trans_F | AAAAGAGCCCAAAGAACAAGGTG | RT-PCR on transgenic soybean roots |
| HgG18_RT-PCR_trans_R | TTTCCGTCCAAGGTCAGTTTTTG | RT-PCR on transgenic soybean roots |
| HgG18i_RT-PCR_loop_F | CGCACTTACAGGCGATTAAAGAG | RT-PCR on transgenic soybean roots |
| HgG18i_RT-PCR_loop_R | CGAAATTCCATACCTGTTCACCG | RT-PCR on transgenic soybean roots |
| GmUBI3_RT-PCR_ref_F | CAACTCTCCACTTGGTGTTGC | RT-PCR on transgenic soybean roots |
| GmUBI3_RT-PCR_ref_R | CCTTCATGACAAGCATGGAAGAC | RT-PCR on transgenic soybean roots |
| HgG18_RNAi_confirm_F | GCGATTTTCTGCGACTGTATGG | qRT-PCR on *H. glycines* for RNAi |
| HgG18_RNAi_confirm_R | GATCGGATCTGAGCCTTCCTG | qRT-PCR on *H. glycines* for RNAi |
| HgActin_qRT-PCR_ref_F | GGCCAACAGAGAAAAGATGACC | qRT-PCR on *H. glycines* for RNAi |
| HgActin_qRT-PCR_ref_R | CAACACGAGACCGGTGGTA | qRT-PCR on *H. glycines* for RNAi |
| 3B05QF | ATTTGATGCACAATTAACACTGCTTGG | qRT-PCR on *H. glycines* for RNAi |
| 3B05QR | CATTGTTTGCTGGTTGGAATATTGTTCGTG | qRT-PCR on *H. glycines* for RNAi |
| 4G06QF | TCCCCATTTCTCAATTGTTCTTCTCCG | qRT-PCR on *H. glycines* for RNAi |
| 4G06QR | CGTCAATGTCTTCACGAAAATTTGCATGC | qRT-PCR on *H. glycines* for RNAi |
| 8H07QF1 | ACAACTGCAGCAACAACAGAATCAGG | qRT-PCR on *H. glycines* for RNAi |
| 8H07QR1 | CTTCCTCGCCATTCATCATCTTGCTC | qRT-PCR on *H. glycines* for RNAi |
| NbPR1a_qRTPCR_F1 | CGACCAGGTAGCAGCCTATG | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NbPR1a_qRTPCR_R1 | TCTCAACAGCCTTAGCAGCC | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NtPR2_qRTPCR_F1 | GGGCTGTTAATTTGCAGTATCC | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NtPR2_qRTPCR_R1 | GGTTTATAACATCTTGGTCTGATGG | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NtWRKY12_qRTPCR_F1 | CTCATCAGCTAGTTCATTTGATGC | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NtWRKY12_qRTPCR_R1 | AGCTCGGTCTTTGTTCTAAAAGC | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NtPI1_qRTPCR_F1 | CTTCAAAGACTATGGTGAAGTTTGC | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NtPI1_qRTPCR_R1 | CAGACTGAGACACATCAAGTTGC | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NbActin1_qRTPCR_F1 | GTTGCTATACAAGCTGTTCTCTCG | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| NbActin1_qRTPCR_R1 | GTCAAGACGAAGAATGACATGTGG | qRT-PCR on N. benth for quantitative PTI and ETI analyses |
| HgG18_full-length_F | ATGGCCATTCTGCTGAAGTGTGT | Molecular Cloning of G18 CDSs |
| mat_HgG18_startF | ATGGACCCCGGCAAAAAAGGAAAGAGC | Molecular Cloning of G18 CDSs |
| HgG18_stopR | TCATTTCCGTCCAAGGTCAGTTTTTGTTGC | Molecular Cloning of G18 CDSs |
| HgG18_geneseq_F1 | ACTCCGAAACGTGGCGGTAA | Primer walking *HgG18* gene clones |
| HgG18_geneseq_R1 | CAACTCTTGTGGGTCGTCCA | Primer walking *HgG18* gene clones |
| HgG18_geneseq_F2 | CGTAATTACGAAGTTAGAGA | Primer walking *HgG18* gene clones |
| HgG18_geneseq_R2 | TGAAATTTAAATATAATTCTACT | Primer walking *HgG18* gene clones |

Table S4 continued

| | | |
|---|---|---|
| HgG18_DELscharged_R | TATTACCTGCAATTCCATTCACCTTGTTCTTT | Site-directed mutagensis, HgG18_Δsupercharged |
| HgG18_DELscharged_F | GAATGGAATTGCAGGTAATACGACTGACGCGA | Site-directed mutagensis, HgG18_Δsupercharged |
| PfiCSPF | GGGCATAATGTAGATCTCTCCAGGGCA | PfiCSP complementation experiments |
| PfiCSP_RI,RR,RII+_pENTR_F | CACCGTAGAACCAAAAAAGCCACATGAAAATAAGCTG | PfiCSP complementation experiments |
| PfiCSP_RI,RR,RII+_OEPCR_R | CTGCATGGGCCTTCCTGATTTGCTCCTGGGGCT | PfiCSP complementation experiments |
| PfiCSP_RI,RR,RII+_OEPCR_F | AGGAAGGCCCATGCAGTGTAACCTGTGGAAAGGGT | PfiCSP complementation experiments |
| PfiCSP_nostopR | GCACTTATCCATTGTACAAACTTCTGCCT | PfiCSP complementation experiments |
| PfiCSP_RI,RR,RII+_stopR | TCAATTCAGAGTAAGCTTATCTGGTTTTTTGTCACC | PfiCSP complementation experiments |
| PfiCSP-HgG18_OEPCR_R | GGGCTATTATTCAGAGTAAGCTTATCTGGTTTTTTGTCACC | PfiCSP complementation experiments |
| PfiCSP-HgG18_OEPCR_F | CTCTGAATAATAGCCCGAGAAAGCCCAACAACG | PfiCSP complementation experiments |
| HgG18_s-charged_stopR | TCATGTTGATTTCGTTCGCAGGAAACGG | PfiCSP complementation experiments |
| GPcontrol-HgG18_pENTR_F | CACCAACCGACGAACTAGTCTGTACCCGAT | PfiCSP complementation experiments |
| GPcontrol-HgG18_OEPCR_R | GGGCTATTTTCCTTGGTCACGCCAATGTCATTGT | PfiCSP complementation experiments |
| GPcontrol-HgG18_OEPCR_F | CCAAGGAAAATAGCCCGAGAAAGCCCAACAACG | PfiCSP complementation experiments |
| HgG18_RNAisense_F | TAAGCA**GGCGCGCC**GAGCAAAGATCCGATCCCAATCCCG | Construction of *PGmUBI:HgG18i* |
| HgG18_RNAisense_R | TAAGCA**ATTTAAAT**TCATTTCCGTCCAAGGTCAGTTTTTGTTGC | Construction of *PGmUBI:HgG18i* |
| HgG18_RNAianti_F | TAAGCA**CCTAGG**GAGCAAAGATCCGATCCCAATCCCG | Construction of *PGmUBI:HgG18i* |
| HgG18_RNAianti_R | TAAGCA**GGATCC**TCATTTCCGTCCAAGGTCAGTTTTTGTTGC | Construction of *PGmUBI:HgG18i* |
| PGmUBI_seq_F | CAGATCCGTACTTAAGATTACGTAATGG | Construction of *PGmUBI:HgG18i* |
| GUSlink_seq_R | CGGTTTGTGGTTAATCAGGAACTGT | Construction of *PGmUBI:HgG18i* |
| GUSlink_seq_F | GGAATTTCGCCGATTTTGCGAC | Construction of *PGmUBI:HgG18i* |
| rbSterm_seq_R | AAACACAGTAAATTACAAGCACAACA | Construction of *PGmUBI:HgG18i* |
| HgG18_XPRESS_F | TAAGCA**ATTTAAAT**ATGGACCCCGGCAAAAAAGGAAAGAGC | Construction of *PGmUBI:HgG18* |
| HgG18_XPRESS_R | TAAGCA**GGATCC**TCATTTCCGTCCAAGGTCAGTTTTTGTTGC | Construction of *PGmUBI:HgG18* |
| supercharged_XPRESS_F | TAAGCA**ATTTAAAT**AATAGCCCGAGAAAGCCCAACAAC | Construction of *PGmUBI:HgG18 mutants* |
| supercharged_XPRESS_R | TAAGCA**GGATCC**TCATGTTGATTTCGTTCGCAGGAAACGG | Construction of *PGmUBI:HgG18 mutants* |
| Nterm_XPRESS_R | TAAGCA**GGATCC**TCAAATTCCATTCACCTTGTTCTTTGGGCTC | Construction of *PGmUBI:HgG18 mutants* |
| Cterm_XPRESS_F | TAAGCA**ATTTAAAT**GCAGGTAATACGACTGACGCGAC | Construction of *PGmUBI:HgG18 mutants* |
| HgG18_mat_pENTR_F | CACCATGGACCCCGGCAAAAAAGGAAAG | Subcloning into pENTR/D-TOPO |
| HgG18_scharged_pENTR_F | CACCAATAGCCCGAGAAAGCCCAACAAC | Subcloning into pENTR/D-TOPO |
| HgG18_DELNterm_F | CACCAATAGCCCGAGAAAGCCCAACAAC | Subcloning into pENTR/D-TOPO |
| HgG18_Nterm_R | TCAAATTCCATTCACCTTGTTCTTTGGGCTC | Subcloning into pENTR/D-TOPO |
| HgG18_Cterm_pENTR_F | CACCGCAGGTAATACGACTGACGCGAC | Subcloning into pENTR/D-TOPO |

## APPENDIX E. CHAPTER 6 SUPPLEMENTARY INFORMATION

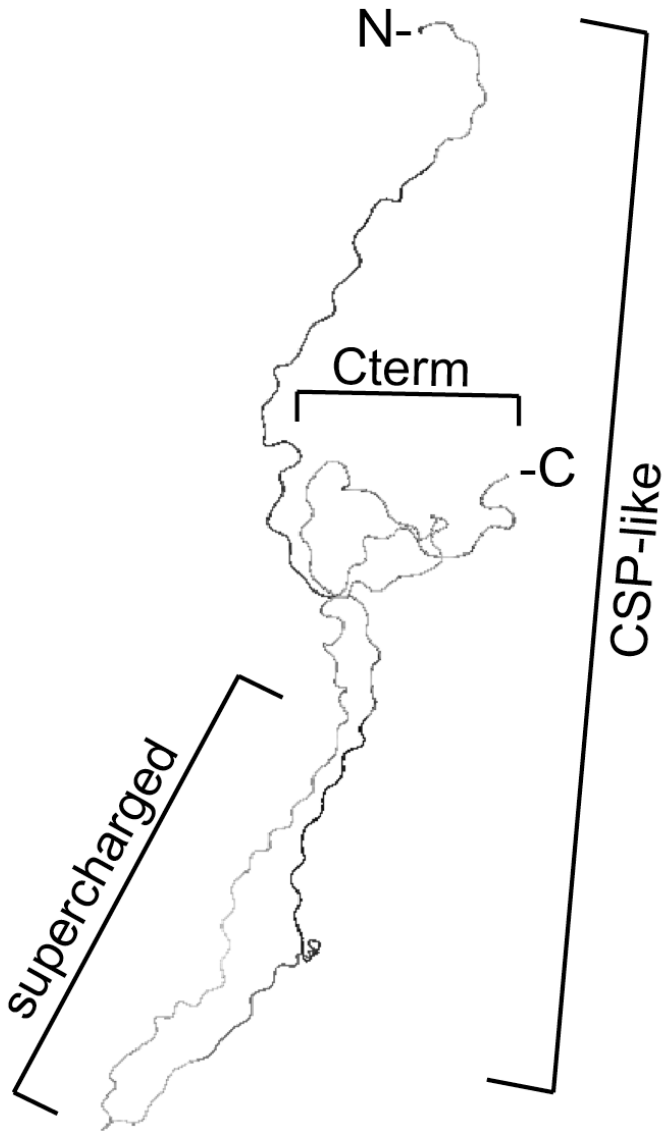



**Fig. S1. *In silico* tertiary structure of HgGLAND18$^{-sp}$.** The tertiary structure was modeled *in silico* for HgGLAND18-3-3 (GenBank: KT954104, isoform 3 with 3 repeats; Noon et al., manuscript submitted) with the signal peptide removed using the I-TASSER server (Roy et al. 2010). Note that aside from numerous coils, the tertiary protein structure was highly disordered. The reliability of the *in silico* tertiary structure was evaluated using the QMEAN server (Benkert et al. 2008) and MolProbity analysis (Chen et al. 2010), both of which suggested that the structure was reliable. The *in silico* tertiary structure was consistent with the protein being highly disordered, which was also suggested from separate secondary structure and disorder predictions.

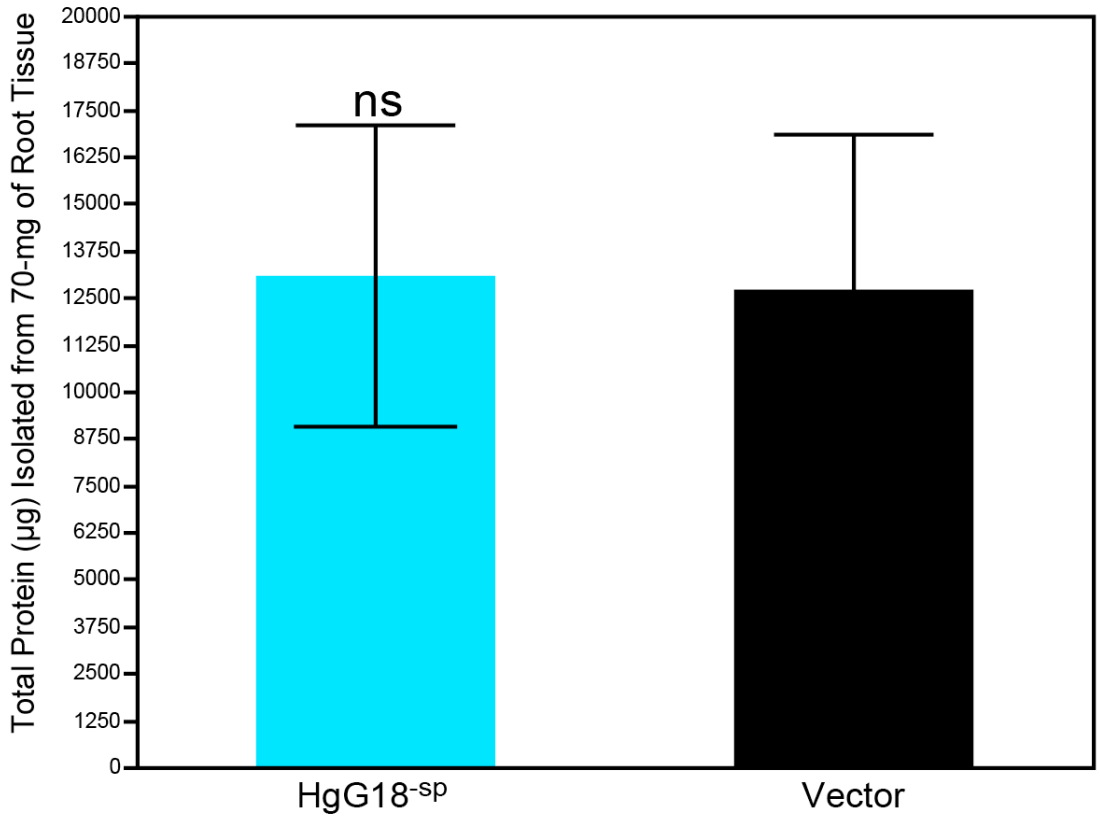



**Fig. S2. Comparison of total protein levels in HgGLAND18(HgG18)$^{-sp}$-expressing and vector control soybean roots.** Data are presented as the means ± one standard deviation (error bars) ($n$ = 8). ns, not significant ($P$ > 0.05). Vector constructs and transgenic soybean roots were generated as in (Noon et al., manuscript submitted).

## Table S1. Complete list of primers used in our study

| Primer Name | Primer Sequence (5'-3') | Application |
|---|---|---|
| NLS1_mut_R | ATCTTTGCTCGCTCCTGCTGCGCCGGGGTCCATAC | Site-directed mutagensis, HgG18 predicted NLS mutants |
| NLS1_mut_F | GTATGGACCCCGGCGCAGCAGGAGCGAGCAAAGAT | Site-directed mutagensis, HgG18 predicted NLS mutants |
| HgG18_NLS1mut_startF | ATGGACCCCGGCGCAGCAGGAGCGA | Site-directed mutagensis, HgG18 predicted NLS mutants |
| NLS2_mut_R | AGAGCTTTTTTATCCGCTGCCGGAGAGCCGTTGT | Site-directed mutagensis, HgG18 predicted NLS mutants |
| NLS2_mut_F | ACAACGGCTCTCCGGCAGCGGATAAAAAAGCTCT | Site-directed mutagensis, HgG18 predicted NLS mutants |
| NLS3_mut_R | ACGGTTTTCTCTTTCTGCTTCTGCTTGCGCTGCTTCTTCTTGAGCTTCCGCTTGTAGAGC | Site-directed mutagensis, HgG18 predicted NLS mutants |
| NLS3_mut_F | GCTCTACAAGCGGAAGCTCAAGAAGAAGCAGCGCAAGCAGAAGCAGAAAGAGAAAACCGT | Site-directed mutagensis, HgG18 predicted NLS mutants |
| HgG18_GUSfusion_F | TAAGCA**AAGCTT**ATGGACCCCGGCAAAAAAGGAAAGAGC | Construction of HgG18:GP for nuclear import experiments |
| HgG18_GUSfusion_R | TAAGCA**CCATGG**TTTTCCGTCCAAGGTCAGTTTTTGTTGC | Construction of HgG18:GP for nuclear import experiments |
| HgG18_mat_pENTR_F | CACCATGGACCCCGGCAAAAAAGGAAAG | Subcloning into pENTR/D-TOPO |
| SV40NLS-GP_pENTR_F | CACCCCGAAGAAGAAGAGAAAGGTGATGGTAGATCTGAGGGTAAATTTCTAGT | Subcloning into pENTR/D-TOPO |
| GUSPlus_pENTR_F | CACCATGGTAGATCTGAGGGTAAATTTCTAGT | Subcloning into pENTR/D-TOPO |
| GUSPlus_nostop_pENTR_R | TACGTGATGGTGATGGTGATGGC | Subcloning into pENTR/D-TOPO |

## APPENDIX F. CHAPTER 7 SUPPLEMENTARY INFORMATION

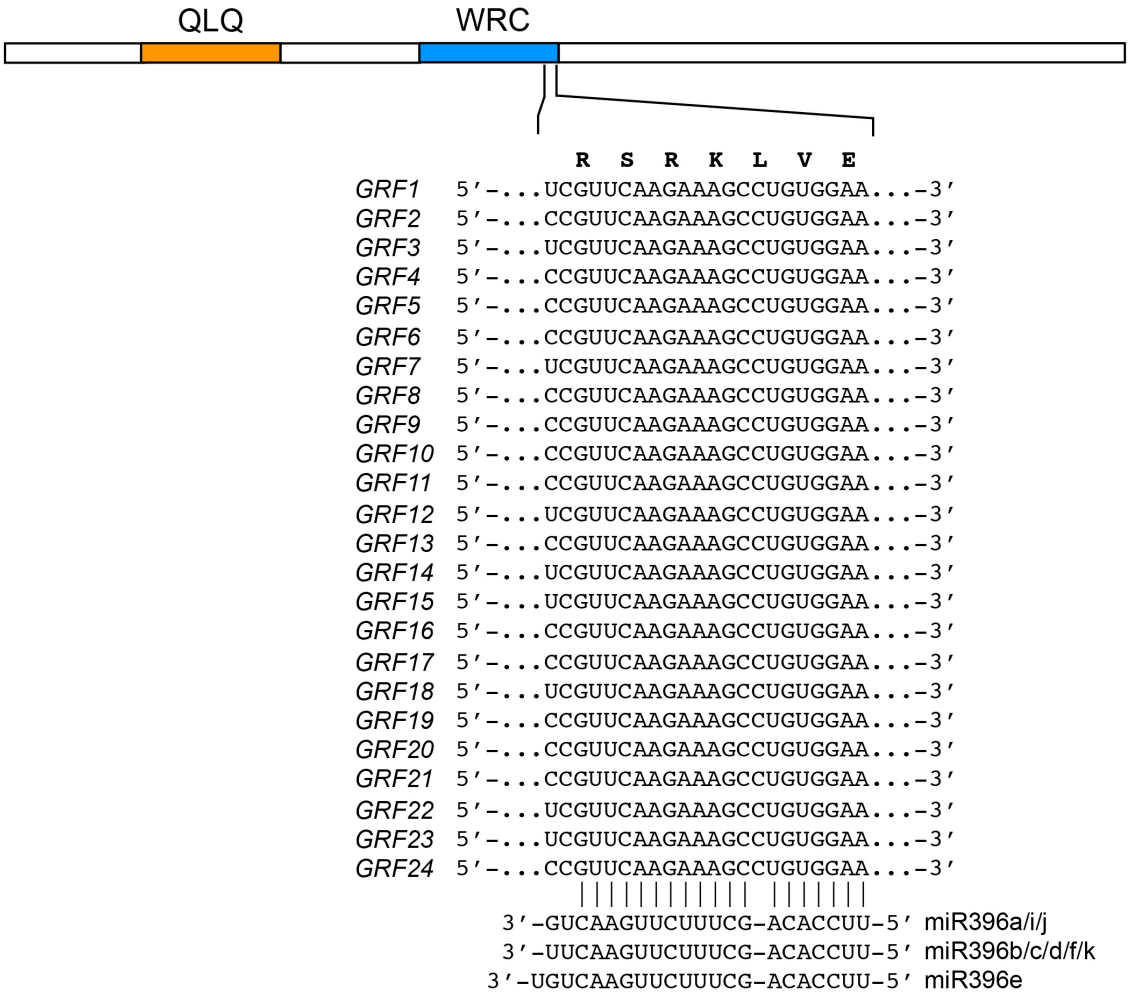



**Fig. S1. Putative miR396 target sites in *GRF1-24*.**

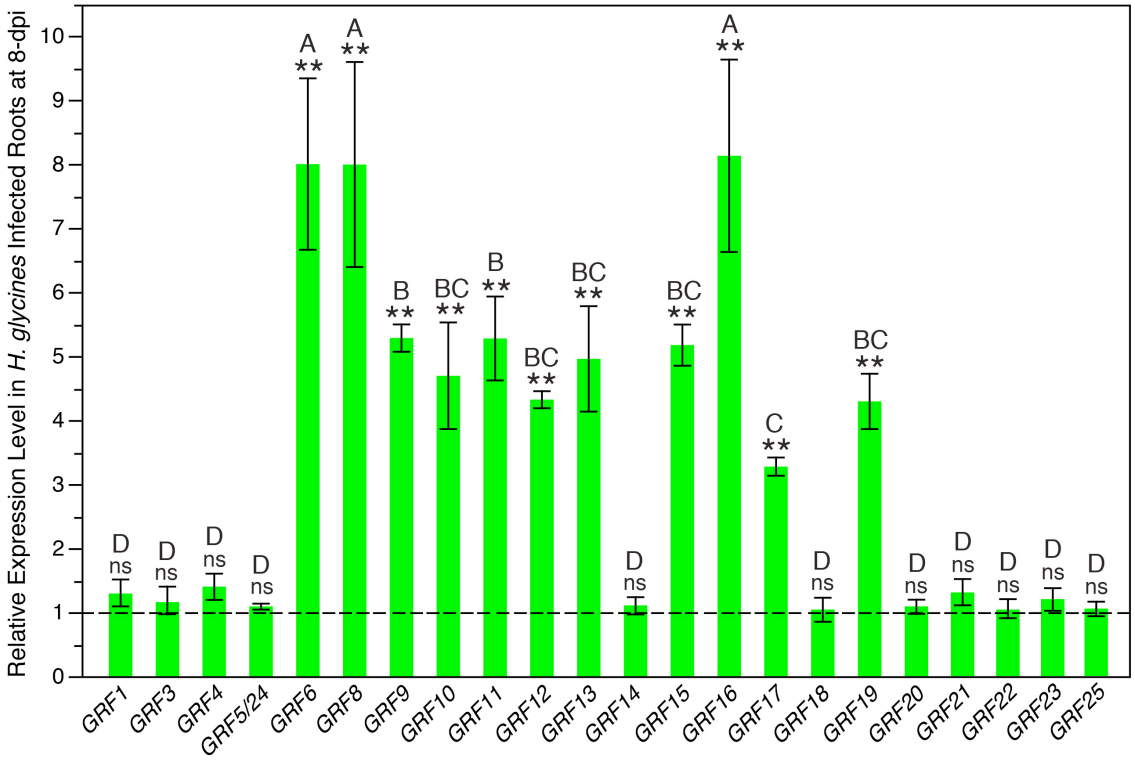



**Fig. S2. qRT-PCR screen of *GRFs* in *H. glycines*-infected roots at 8-dpi.** Expression levels are relative to mock; baseline expression is set to 1.0 and indicated with a dashed line. Error bars represent ± one standard deviation from the mean. For each *GRF*, the bottom label (ns or **) represents statistical comparisons with mock, and the top label represents statistical cross comparisons between the *GRFs* presented as significance groups ($P < 0.05$). **, $P < 0.01$; ns, not significant ($P > 0.05$).

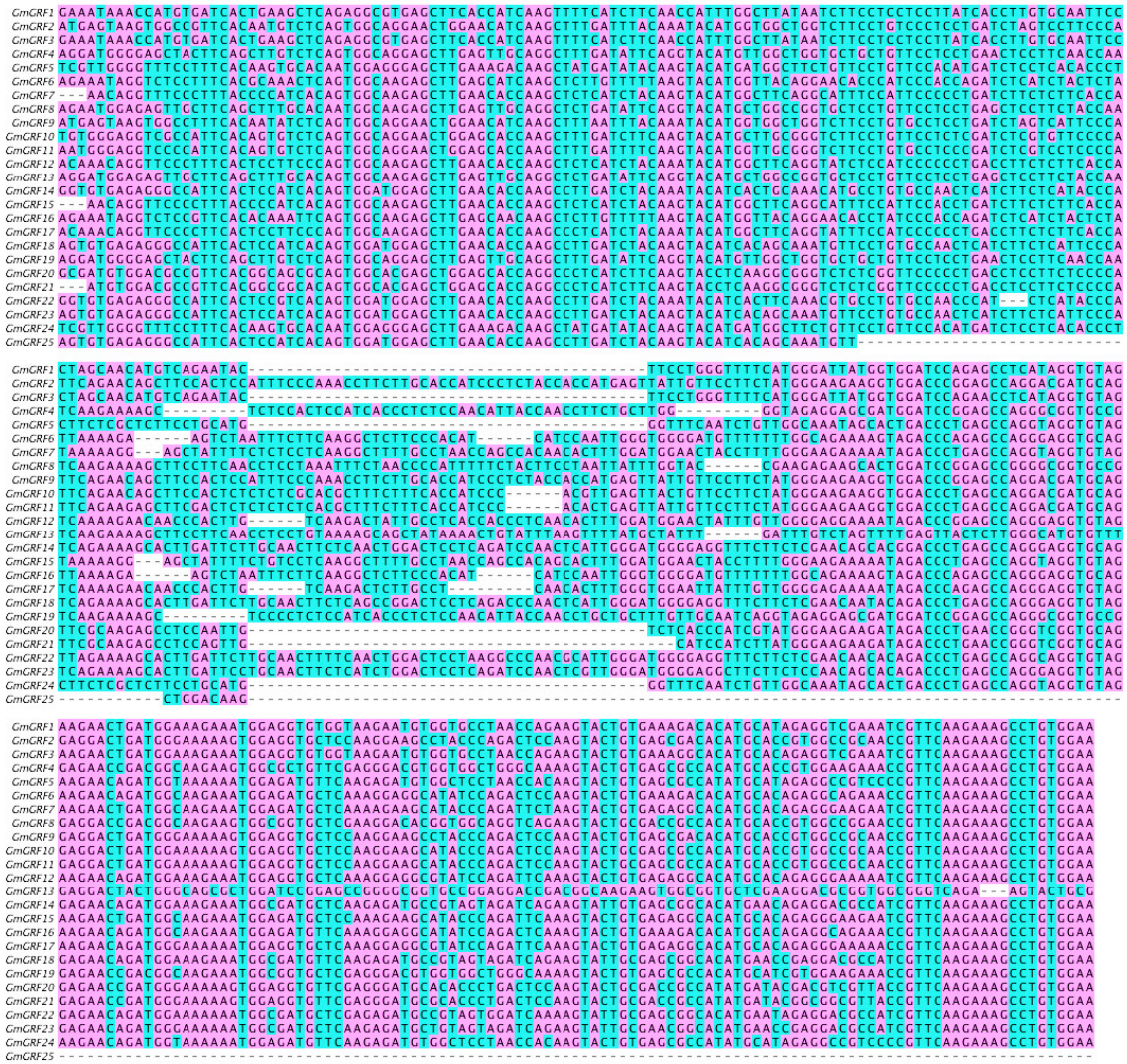



**Fig. S3. Multiple sequence alignment between nucleotides 1-333 from *GRF9* (RNAi target sequence) and *GRF1-25.***

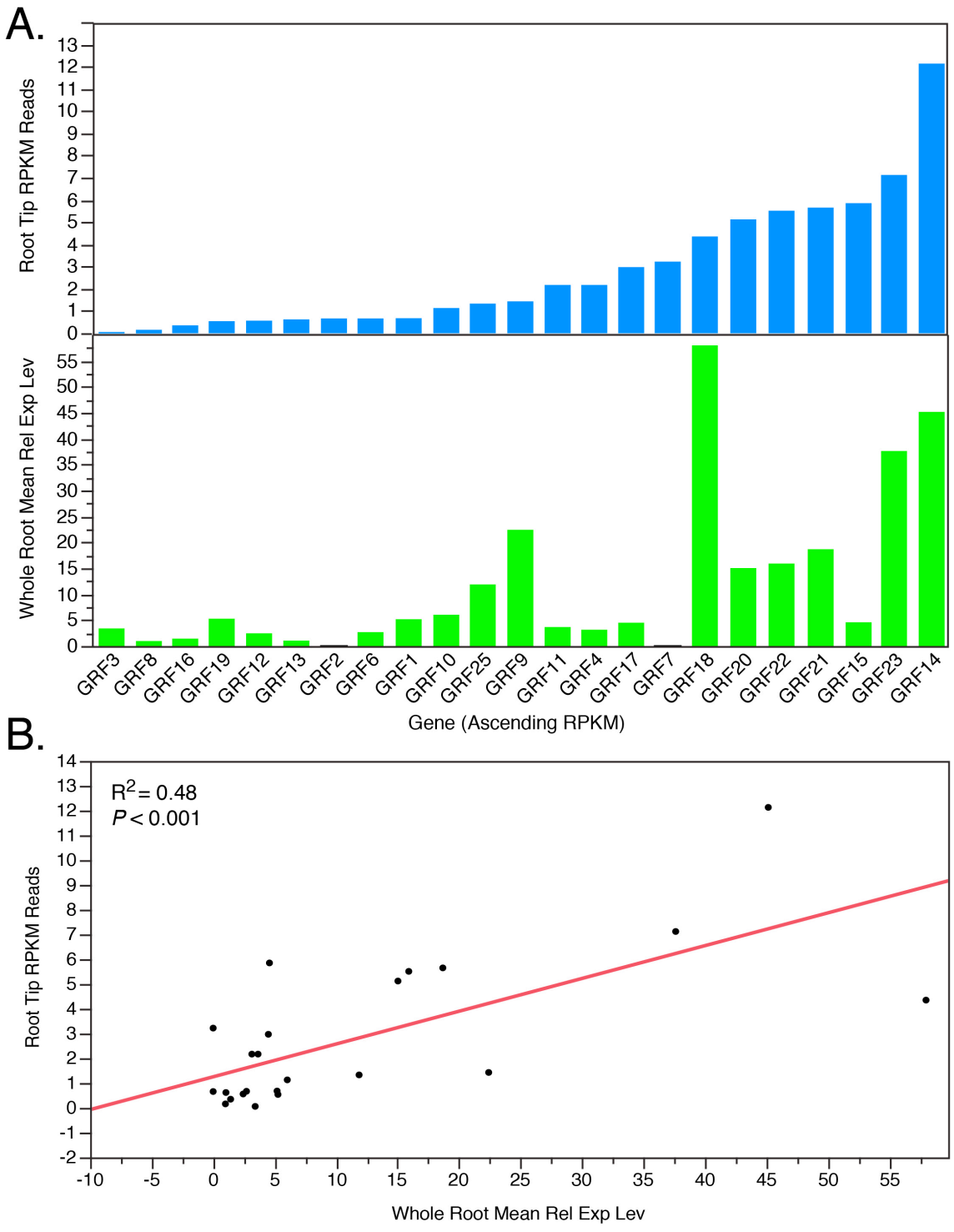



**Fig. S4. Comparison between SFGD root tip RNA-seq data and our whole root qRT-PCR data for native expression of soybean *GRFs*.** RPKM, Reads Per Kilobase of transcript per Million mapped reads.

**Table S1.**

| Gene | Genome Coordinates and Gene Calls from Soybase |
|---|---|
| *MIR396a* | Gm13: 26338134-26338273 [-] |
| *MIR396b* | Gm13: 26329931-26330056 [+] |
| *MIR396c* | Gm13: 43804777-43804893 [+] |
| *MIR396d* | Gm17: 9053049-9053156 [-] |
| *MIR396e* | Gm17: 35366545-35366658 [-] |
| *MIR396f* | Gm15: 556691-556841 [-] |
| *MIR396g* | Gm17: 9053051-9053201 [-] (matches *MIR396d*) |
| *MIR396h* | Gm14: 13971419-13971566 [+] |
| *MIR396i* | Gm17: 9044850-9044984 [+] |
| *MIR396j* | Gm16: 31162190-31162323 [-] |
| *MIR396k* | Gm17: 8782362-8782224 [-] |
| *GRF1* | *Glyma01g34650* |
| *GRF2* | *Glyma01g44470* |
| *GRF3* | *Glyma03g02500* |
| *GRF4* | *Glyma03g35010* |
| *GRF5* | *Glyma04g40880* (nearly identical to *GRF24*) |
| *GRF6* | *Glyma07g04290* |
| *GRF7* | *Glyma09g07990* |
| *GRF8* | *Glyma10g07790* |
| *GRF9* | *Glyma11g01060* |
| *GRF10* | *Glyma11g11820* |
| *GRF11* | *Glyma12g01730* |
| *GRF12* | *Glyma13g16920* |
| *GRF13* | *Glyma13g21630* |
| *GRF14* | *Glyma14g10090* |
| *GRF15* | *Glyma15g19460* |
| *GRF16* | *Glyma16g00970* |
| *GRF17* | *Glyma17g05800* |
| *GRF18* | *Glyma17g35090* |
| *GRF19* | *Glyma19g37740* |
| *GRF20* | *Glyma09g34560* |
| *GRF21* | *Glyma01g35140* |
| *GRF22* | *Glyma14g10100* |
| *GRF23* | *Glyma17g35100* |
| *GRF24* | *Glyma06g13960* (nearly identical to *GRF5*) |
| *GRF25* | *Glyma08g34500* |

**Table S2.**

| Primer ID | Primer Sequence (5'-3') |
|---|---|
| pre-miR396a/i_qPCRF | CTGCATCCAAAGAGTTCCTTTGC |
| pre-miR396b/d/k_qPCRF | AAGTCCTGGTCATGCTTTTCCAC |
| pre-miR396c/f_qPCRF | GTCCTGTTATGCTTTTCCACAGC |
| pre-miR396e/h/j_qPCRF | CTTGAACTGTGTTGTGAGGCTTC |
| miR396a/i/j_qPCRF | TTCCACAGCTTTCTTGAACTGAA |
| miR396b/c/d/f/k_qPCRF | TTCCACAGCTTTCTTGAACTTAA |
| miR396e_qPCRF | TTCCACAGCTTTCTTGAACTGTAA |
| GRF1_qPCRF | ATGAACCGGCAGTTACCATTGCG |
| GRF1_qPCRR | ATGGTTCACTTGCTGGGACTGAC |
| GRF2_qPCRF | TGACTGTAACTGCTGGCAGCAG |
| GRF2_qPCRR | CTGATCCAGAACCAGTCCCTTGG |
| GRF3_qPCRF | CAGCAGCTACTGCTACCAGTTCC |
| GRF3_qPCRR | ATGAATGCTTCCCTGGCTTGGAC |
| GRF4_qPCRF | CTTCGATGGGAGGTCCACTTGC |
| GRF4_qPCRR | TCCACGAGGCAAGTGACGATG |
| GRF5/24_qPCRF | ATCTCCATCTGGGGTGTTGCAG |
| GRF5/24_qPCRR | AGCAAGGCCATCTCAGAATTGGC |
| GRF6_qPCRF | CATGACACCACCCTCTTCCTTCC |
| GRF6_qPCRR | GAAAAGGGTTGGGACTCAATAGGC |
| GRF7_qPCRF | AACAAACACAAACTCAAATGCTTC |
| GRF7_qPCRR | CACCAACACCACCCTCAACACTA |
| GRF8_qPCRF | TGGGGATGGCAAATCAAATGGCC |
| GRF8_qPCRR | TCAACGACACATCCGAGGGATCG |
| GRF9_qPCRF | CACTTTGATGCCAACCGGTGGAG |
| GRF9_qPCRR | GGACTGACCCTCTTGCTTCACAG |
| GRF10_qPCRF | CTCACTGTCACTGGGGGTAGTGG |
| GRF10_qPCRR | CAGTTCCAGAATCGGTACCCTGG |
| GRF11_qPCRF | CTGAGGGTGGTGAACACTGCTTC |
| GRF11_qPCRR | AGAGGCAACTCTGGTTGACATCG |
| GRF12_qPCRF | GCAGCAGTGGCAACAAACACAG |
| GRF12_qPCRR | GGAGAGTGCATGTGCAGGACTG |
| GRF13_qPCRF | GTGGATGGGGATGGAAATTCAGC |
| GRF13_qPCRR | TCAACGACACATCCGAGGAAGTG |
| GRF14_qPCRF | CACAACAAAGTGCACCACCCTC |
| GRF14_qPCRR | GTAGCATGGGAAGGGCAGAGC |
| GRF15_qPCRF | AATAGCACTCTCTCACCTCTTGC |
| GRF15_qPCRR | GAGGATAGTGGTGGTGTTGGTGAG |
| GRF16_qPCRF | CCACCACCACCCTCTTCTTTCCC |
| GRF16_qPCRR | TGGGAAAAGGGTTGGGACTGATC |
| GRF17_qPCRF | CAGTGCCCCCTTGTTTCTCGAC |
| GRF17_qPCRR | CCATACACATACCTGCAGTCAGTG |
| GRF18_qPCRF | TAGGTCATCCCCAACTGGGGTG |
| GRF18_qPCRR | CATCTCTGCTCCCAGCACTGC |
| GRF19_qPCRF | AAAACAATGCTGCTGGCCCTAGC |

Table S2 continued

| | |
|---|---|
| GRF19_qPCRR | GGAAGCAGCAGCAGCATTTCCG |
| GRF20_qPCRF | GTTGCTAATGTTTCTGCAACTGCTGC |
| GRF20_qPCRR | GTGTTGGTGTTGGTGTTGGTGC |
| GRF21_qPCRF | TGTTTCTGGAACTGCAACTGCTGC |
| GRF21_qPCRR | CCCTTGCACCATTGGTGTGC |
| GRF22_qPCRF | GCCGGTGGAAGGTCTTTTCGTGG |
| GRF22_qPCRR | TGGCCCTCTCACACCAGCCA |
| GRF23_qPCRF | AGGCCATGCCCTCACCACCA |
| GRF23_qPCRR | TGTTGTCCTTTGGTTTCAGCACGGT |
| GRF25_qPCRF | TCCAAAGTCAGAGACTTTATTGGTAGAC |
| GRF25_qPCRR | TGCTTATGCTTCCTGTATTTTTGCTGTA |
| GmUBQ3_qPCRF | GACCAGCAACGTCTCATTTTCGC |
| GmUBQ3_qPCRR | GTGTCGGAGCTTTCAACCTCTAGG |
| GRF6_5'-RACE_outer | TTCCATGAACATGCCTGCCA |
| GRF6_5'-RACE_inner | CACATTCAACTCGGTTAAAGACCA |
| GRF8_5'-RACE_outer | GTCCTCTCCATAGCCAGTGGTCA |
| GRF8_5'-RACE_inner | TCACGGTTTCTAGACAAGCTCTTGCTG |
| GRF9_5'-RACE_outer | AGAGCTAGTTGCAGAGAAATTTGACGAT |
| GRF9_5'-RACE_inner | TGGGTTTGGTCTGTTCCAGAAGCAG |
| GRF10_5'-RACE_outer | CACATGTCAGACCATCACCCA |
| GRF10_5'-RACE_inner | TGTGGTCAGTTCCAGAATCGG |
| GRF11_5'-RACE_outer | AAATTGAATCTCATTCTCACCATGT |
| GRF11_5'-RACE_inner | GAAGGCCATGTGTTTTCCAGC |
| GRF12_5'-RACE_outer | CTCTGAAGTTGCAAGTAGGAGTACTCGT |
| GRF12_5'-RACE_inner | TACTTCATCAAGCCAAGCAGGCTTG |
| GRF13_5'-RACE_outer | TCCTGAGTTGTTGTTCTGGCAACCA |
| GRF13_5'-RACE_inner | CCATCCCCATCCACATGATCATGGT |
| GRF15_5'-RACE_outer | TGTCATCCAAATCAAGCCAAGAGCCT |
| GRF15_5'-RACE_inner | GACAAGCCAAAGGAACTCCTCTGCT |
| GRF16_5'-RACE_outer | GGAGCTCATTGATAGTTGTTGGTACGAG |
| GRF16_5'-RACE_inner | CATGAGAATAAGAGCCAGCAGAGTCCA |
| GRF17_5'-RACE_outer | TTGCCACTGAGGCTCTGAAG |
| GRF17_5'-RACE_inner | AAGGCATGCTCATCCACCTC |
| GRF19_5'-RACE_outer | GGAAGCAGCAGCAGCATTTCCG |
| GRF19_5'-RACE_inner | AGATCAAATGGGGATTTGAGTGAAGCAG |
| PMIR396a_CLONE_F | TAAGCA**CTGCAG**ATCAAATCTCAGCTGGCCTTC |
| PMIR396a_CLONE_R | TAAGCA**GTCGAC**TATTGAAGACTGAATATACCC CACAAAC |
| PMIR396b_CLONE_F | TAAGCA**CTGCAG**ACAGTGAATAAAGGAAGAGA GGCAGT |
| PMIR396b_CLONE_R | TAAGCA**GTCGAC**AGAGGAAAATTGAGCTCAGA TCTCTCG |
| PMIR396c_CLONE_F | ATTCGT**CTGCAG**GCATCCTTTGCTGCAACTTCAA CTC |

Table S2 contintued

| | |
|---|---|
| PMIR396c_CLONE_R | TAAGCA**GTCGAC**AAGAAAGCATGGAAGTTGAAATCCTGAAGC |
| PMIR396d_CLONE_F | ATTCGT**CTGCAG**ATTGCCTTTTTCCGCTCATGTTACCAT |
| PMIR396d_CLONE_R | TAAGCA**GTCGAC**AGGACTTGAGAGGAAAATTGAGCTCAG |
| PMIR396e_CLONE_F | TAAGCA**CTGCAG**GGATGGTGTCGATGAAGACGTGC |
| PMIR396e_CLONE_R | TAAGCA**CCATGG**GAAAAAGACCATTCCCAAAGCAACACCCT |
| PMIR396f_CLONE_F | TAAGCA**CTGCAG**GAGACCGCTAATATGGCTCCGAATG |
| PMIR396f_CLONE_R | TAAGCA**GTCGAC**AATATATAGCAGAATGAATTGAGTTAGAGGAGTGGAAAGGAG |
| PMIR396i_CLONE_F | TAAGCA**CTGCAG**ACTCTAATGTGATGAGATACACGCGTTC |
| PMIR396i_CLONE_R | TAAGCA**GTCGAC**TGATATTGAAGACTGAAGATAAATACTCCACAAGCCC |
| PMIR396j_CLONE_F | TAAGCA**CTGCAG**CTAGAGGAGGAAGTGGTAGCGATCGA |
| PMIR396j_CLONE_R | TAAGCA**CCATGG**ATTCCCAAAGCAACACCCTACGGA |
| PMIR396k_CLONE_F | TAAGCA**CTGCAG**GAAATTATTGAAACACACGCTTATTGCC |
| PMIR396k_CLONE_R | TAAGCA**GTCGAC**AATTGAGCTCAGATCTCACACACACAG |
| PGRF6_CLONE_F | TAAGCA**GTCGAC**GACTAAAGAGGCTCCCAGGTGA |
| PGRF6_CLONE_R | TAAGCA**CCATGG**TTTGAGACAACAAAAGGAAGGAAGAGTG |
| PGRF8_CLONE_F | TAAGCA**AAGCTT**CAGCACAGTGGCACAGTGTAGG |
| PGRF8_CLONE_R | TAAGCA**TCTAGA**GGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAA |
| PGRF9_CLONE_F | TAAGCA**CTGCAG**AGAGGTGTGGCTTGCAAGGA |
| PGRF9_CLONE_R | TAAGCA**GTCGAC**TCTCCTCCCCCACTCTTCTCTTCTC |
| PGRF10_CLONE_F | TAAGCA**CTGCAG**GTATACTATGTCACTATGCCACTGGT |
| PGRF10_CLONE_R | TAAGCA**GTCGAC**GTTTTCCACTACTCTTACTTAGTCCTAT |
| PGRF11_CLONE_F | TAAGCA**CTGCAG**AGGGTGTAGTCCATTGTACCGTTCTAGA |

Table S2 continued

| | |
|---|---|
| PGRF11_CLONE_R | TAAGCA**GTCGAC**GTTTTCCACTGTTGTCTACTTG CGCTAC |
| PGRF12_CLONE_F | TAAGCA**CTGCAG**GGATGCTGTTGCTAACATATG CTGCT |
| PGRF12_CLONE_R | TAAGCA**GTCGAC**TATCGAGCTGAGGGGAAAGG GAAAGG |
| PGRF13_CLONE_F | TAAGCA**CTGCAG**CACAGAATGACAGCAGAGTG GCA |
| PGRF13_CLONE_R | TAAGCA**GTCGAC**GGAGAGAGAGAGAGAAGAGC TATTGAGC |
| PGRF15_CLONE_F | TAAGCA**GTCGAC**AGGTAAAGGAACAGTAAAAG CTCCAAGC |
| PGRF15_CLONE_R | TAAGCA**CCATGG**CATCAAAGTAAGTAGAGAAG TAGAAGAGGAAAATAGAAGTGC |
| PGRF16_CLONE_F | TAAGCA**GTCGAC**GTTCTTAATTGGGTGGTCTAC CCAGCTA |
| PGRF16_CLONE_R | TAAGCA**CCATGG**TTTGAGACAACAAAAAGAAG AAAGGGTTTGTGTTTG |
| PGRF17_CLONE_F | TAAGCA**CTGCAG**TAATCCGTTCTTCAACTGTTG GTGAAGA |
| PGRF17_CLONE_R | TAAGCA**GTCGAC**CGAGCAGAGGGGAAAGGGAA AGA |
| PGRF19_CLONE_F | TAAGCA**CTGCAG**CAACCGCTTGATTTAATCATC GCTGGTC |
| PGRF19_CLONE_R | TAAGCA**GTCGAC**GGAGAGAGAGAGAGAGAGAG AGAGAGAGAGAGAGAGAGACTTG |
| pre-miR396a_CLONE_F | TAAGCA**ATTTAAAT**TTGTGGGGTATATTCAGTCT TCAATA |
| pre-miR396a_CLONE_R | TAAGCA**GGATCC**TGGAGTTTGAGATAAGCTAGT CCGTT |
| pre-miR396b_CLONE_F | TAAGCA**ATTTAAAT**GAGAGATCTGAGCTCAATT TTCCTCT |
| pre-miR396b_CLONE_R | TAAGCA**GGATCC**GTGATGGTACCAAATGAAAG CAATTAAA |
| pre-miR396c_CLONE_F | TAAGCA**ATTTAAAT**CAGGATTTCAACTTCCATG CTTTCTT |
| pre-miR396c_CLONE_R | TAAGCA**GGATCC**TTGTGTGGGATTTTTTAGAAC CAATT |
| pre-miR396d_CLONE_F | TAAGCA**ATTTAAAT**TGAGCTCAATTTTCCTCTCA AGTCCT |
| pre-miR396d_CLONE_R | TAAGCA**GGATCC**ACCAAATTAAGGCAATTAAA AGCCTGAA |
| pre-miR396e_CLONE_F | TAAGCA**ATTTAAAT**GTGTTGCTTTGGGAATGGT CTTTTTC |

Table S2 continued

| | |
|---|---|
| pre-miR396e_CLONE_R | TAAGCA**GGATCC**AGGAATTGATGAACCAGATC ATCTCA |
| pre-miR396f_CLONE_F | TAAGCA**ATTTAAAT**CACTCCTCTAACTCAATTCA TTCTGC |
| pre-miR396f_CLONE_R | TAAGCA**GGATCC**GGTGTACTGTGTAAGTTCTTA TAACCA |
| pre-miR396i_CLONE_F | TAAGCA**ATTTAAAT**TTGTGGAGTATTTATCTTCA GTCTTCA |
| pre-miR396i_CLONE_R | TAAGCA**GGATCC**TTTGAGATAAGCTAGTTCGTT GTTGT |
| pre-miR396j_CLONE_F | TAAGCA**ATTTAAAT**ATTCCGTAGGGTGTTGCTTT GGGA |
| pre-miR396j_CLONE_R | TAAGCA**GGATCC**TCAAAGAGGAATTGATGAAC CAGATCATCTC |
| GRF9_CLONE_startF | TAAGCA**GGCGCGCC**ATGAGTAAGTGGCCTTTCA CAATATCTC |
| GRF9_CLONE_stopR | TAAGCA**CCTAGG**TTAAATCTCACCATGTGGGGA ATGAG |
| rGRF9_OEPCR_R | CTCTACGGGTTTCCTAGAGCGGTTGCGGCCACG GTGCATGTGTCG |
| rGRF9_OEPCR_F | CCGCTCTAGGAAACCCGTAGAGTCACAAACTAT GACACAGTCATCATC |
| GRF9_RNAi_sense_F | TAAGCA**ATTTAAAT**ATGAGTAAGTGGCCTTTCA CAATATCTC |
| GRF9_RNAi_sense_R | TAAGCA**ATTTAAAT**TTCCACAGGCTTTCTTGAAC GG |
| GRF9_RNAi_anti_F | TAAGCA**GGATCC**ATGAGTAAGTGGCCTTTCACA ATATCTC |
| GRF9_RNAi_anti_R | TAAGCA**GGATCC**TTCCACAGGCTTTCTTGAACG G |

# APPENDIX G. ANALYSIS OF TRANSGENIC SOYBEANS WITH NEMATODE-INDUCIBLE EXPRESSION OF MIR396 IN SYNCYTIA

In chapter 6, we determined using soybean hairy roots that overexpression of all functional pre-miR396 (i.e., pre-miR396 whose predicted stem-loop structures placed the miR396/miR396* duplex within the stems rather than overlapping with the loops such as pre-miR396h) substantially reduces susceptibility to the soybean cyst nematode, *Heterodera glycines*. Additionally, we showed that both RNA interference (RNAi) against *GRFs* and overexpression of a miR396-resistant *GRF9* mutant phenocopies pre-miR396 overexpression. Thus, this system (called the miR396-*GRF6/8-13/15-17/19* regulatory network) represents an attractive target for engineering transgenic, synthetic resistance to *H. glycines* in soybean.

In 2012 (i.e., well before experiments were conducted in soybean hairy roots), with the prior expectation that we would obtain results comparable to those described above, we began a project to generate stable transgenic soybeans (*Glycine max*) that we hoped through overexpression of pre-miR396 would interfere with syncytium formation, and thus prevent or at least negatively affect *H. glycines* development to the adult female stage. A technical challenge for this project was that the miR396-*GRF* regulatory network, at that time, was expected to be essential for plant development. Thus, interfering with this network throughout the soybean plant would likely have caused extensive defects in growth and development.

Kandoth et al. (2011) documented promoters for multiple soybean genes that showed specific activities within *H. glycines* syncytia in soybean hairy roots, and that

were already known to have negligible activity throughout the rest of the soybean plant. Thus, these soybean promoters represented ideal candidates for nematode-inducible expression in *H. glycines* syncytia. Accordingly, one of these promoters—*non-race-specific-disease resistance gene 1* (***N****DR1*)/*harpin-induced-1* (***H****IN1*)-***L****ike-**1*** (*PGmNHL1*)—was selected by us in order to overexpress pre-miR396 specifically within syncytia. However, in (Kandoth et al. 2011), *PGmNHL1*, as well as the other candidate nematode-inducible promoters, were studied in a resistant soybean cultivar. As we were interested in generating stable transgenic soybean plants using a nonresistant cultivar so that we could avoid the yield reductions caused by resistance, I had to test whether *PGmNHL1* also showed specific activity within *H. glycines* syncytia of cultivar (cv.) Williams 82, which is highly susceptible. Using GUSPlus as a reporter, I found that *PGmNHL1* was active specifically within syncytia at all *H. glycines* stages evaluated (J2-J4) in Williams 82 hairy roots (e.g., see Fig. 1).

**Fig. 1. *PGmNHL1* is active specifically within syncytia in *H. glycines*-infected soybean Williams 82 hairy roots.** Image shows parasitic *H. glycines* J2s feeding from

blue syncytia (blue indicates *PGmNHL1* activity via GUSPlus staining). Scale bar equals one millimeter.

Since at the time of starting this project, only pre-miR396a, b, c, d, e, f, and i had been reported to miRBase, I only attempted to use these seven functional pre-miR396 for the transgenics; gma-miR396j was not reported to miRBase until 2014. All seven pre-miR396 were cloned into pGEM-T Easy and sent to Wayne Parrott's lab at the University of Georgia. Peter LaFayette generated the *PGmNHL1:pre-miR396:rbcSterm* vectors for all seven pre-miR396. The vectors for pre-miR396b, c and i were sent to Lila Vodkin's lab at the University of Illinois, pre-miR396d and f to John Finer's lab at Ohio State University, and Wayne Parrott's lab took responsibility of pre-miR396a and e. For soybean transformation, all groups used particle bombardment of embryonic soybean tissue (see http://www2.oardc.ohio-state.edu/plantranslab/d20.htm for extensive details on the procedure). Unfortunately, T1 seed was unable to be obtained for pre-miR396a, c, and i, but T1 seed was successfully obtained for pre-miR396b, d, e and f. Table 1 shows all T1 seed that was obtained from the three soybean transformation groups by January 13, 2015. It is important to note that due to miscommunication between the transformation groups, the Parrot lab generated transgenic seed using soybean cv. Williams 82, the Finer lab used cv. JackX and the Vodkin lab used cv. Jx4. All transgenic soybean seed was stored in plastic bags inside of steel containers at ~4˚C.

Before phenotyping, we genotyped some of the T1 seed (Table 1) to ensure successful transformation (with an expected 3:1 ratio of transgenic:nullizygous seed). We also performed RT-PCR analysis on a few *H. glycines*-infected plants that were predetermined to be positive for the transgene. Because of this experimental design, we

had to process *H. glycines*-infected root tissue for both genotyping and for testing transgene expression under nematode infection.

Soybean seeds were planted into a sand-soil mixture in cone-tainers, were grown in a growth chamber for one week, and then were inoculated with 5,000 *H. glycines* line PA3 eggs.

**Table 1. Transgenic soybean seed for pre-miR396 overexpression constructs**

| OE Construct | Trans. Group | Event | Plants (or lines)[†¥] | Amt. of Seed |
|---|---|---|---|---|
| pre-miR396b | Vodkin | 1 | 97-2-1-: 1, 4, 7, 8 | >110, >110, 82, 40 |
| pre-miR396b | Vodkin | 2 | 97-2-4-: 1, 2, 4, 5 | >110, >110, >110, >110 |
| pre-miR396b | Vodkin | 3 | 97-4-1-: 1, 4, 5, 10 | 45, 25, 37, 26 |
| pre-miR396b | Vodkin | 4 | 97-4-2-: 1, 3, 9, 13 | >110, >110, 83, 59 |
| pre-miR396b | Vodkin | 5 | 97-4-3-: 4, 6, 7, 9 | >110, >110, >110, 85 |
| pre-miR396b | Vodkin | 6 | 97-4-4-: 2, 3, 5, 6 | >110, >110, >110, >110 |
| pre-miR396b | Vodkin | 7 | 97-4-5-: 2, 4, 6, 7 | >110, 72, 83, >110 |
| pre-miR396b | Vodkin | 8 | 97-4-6-: 1, 2, 3, 4 | 59, 39, 74, 46 |
| pre-miR396e | Parrott | 4 | A*, **B*** | 50, 46 |
| pre-miR396e | Parrott | 10 | **A***, B* | 50, 75 |
| pre-miR396e | Parrott | 21 | B, C* | 82, 62 |
| pre-miR396e | Parrott | 27 | A, B | 52, 39 |
| pre-miR396e | Parrott | 32 | A*, B* | 78, 51 |
| pre-miR396d | Finer | 3 | P1 | 39 |
| pre-miR396d | Finer | 5 | **P1*** | >110 |
| pre-miR396f | Finer | 2 | P1 | 24 |
| pre-miR396f | Finer | 3 | P1*, P2, P3 | >110, >110, 2 |
| pre-miR396f | Finer | 4 | P1, P2 | 109, >110 |
| pre-miR396f | Finer | 5 | P1, P2, P3 | 26, 8, >110 |
| pre-miR396f | Finer | 6 | P2, P3, P4 | 92, >110, >110 |
| pre-miR396f | Finer | 10 | P4, P5*, P6, P8 | >110, >110, >110, 15 |
| pre-miR396f | Finer | 11 | P1, P2, P3, P4 | >110, 32, >110, 57 |

[†] Plants or lines that had 3 or 4 good samples (based on successful amplification of *GmUBI3* reference) genotyped are underlined, and samples determined as positive for the pre-miR396 transgene contain an asterisk.

[¥] Plants or lines that were determined to express the transgene in *H. glycines*-infected soybean root tissue in RT-PCR are indicated with bold font.

*H. glycines*-infected root tissue was harvested one week after inoculation. Genomic DNA was isolated using the ChargeSwitch gDNA Plant Kit (Life Technologies). Initial genomic PCRs were performed on each gDNA sample for the *GmUBI3* reference gene (see chapters 5 and 6) using the following primers: 5'-CAACTCTCCACTTGGTGTTGC-3' and 5'-CCTTCATGACAAGCATGGAAGAC-3'. Samples positive for reference gene amplification were tested for amplification of the respective pre-miR396 transgene (Table 1). Forward primers used for both genotyping and RT-PCR tests for transgene expression were as follows: pre-miR396b_F: 5'-GAGAGATCTGAGCTCAATTTTCCTCT-3'; pre-miR396b_loop_F: 5'-GCATCTTATATCTCTCCACCTCCA-3'; pre-miR396d_F: 5'-TGAGCTCAATTTTCCTCTCAAGTCCT-3'; pre-miR396e_F: 5'-GTGTTGCTTTGGGAATGGTCTTTTTC-3'; pre-miR396f_F: 5'-CACTCCTCTAACTCAATTCATTCTGC-3'; pre-miR396f_loop_F: 5'-ATGCCTAATGCAGCTATTGATGTG-3'. The following primer targeted the terminator sequence and was used as a universal reverse for both genotyping and RT-PCR: rbcSterm_FUE_R: 5'-AGTAAATTACAAGCACAACAAATGGTAC-3'.

Unfortunately for pre-miR396b, we were unable to get specific bands from genomic PCR for the transgene using either pre-miR396b_F or pre-miR396b_loop_F in combination with rbcSterm_FUE_R. However, for pre-miR396d, e and f, transgene amplification was successful and those results are summarized in Table 1; almost all events for pre-miR396e were determined positive for the transgene. Note that the following forward primers resulted in specific amplification for pre-miR396d, e and f transgenes in genomic PCR: pre-miR396d_F, pre-miR396e_F and pre-miR396f_F,

respectively. Thus, overall the genotyping was successful for pre-miR396d, e and f T1 seed, even though not all of the events were proven to be transgenic. For pre-miR396b, new primer combinations will probably need to be tested for genotyping. I suggest testing the following primers originally developed by Peter LaFayette: NHL1-880F: 5'-TTCCCATAGAATCTTCAAAGCTTCC-3' and RbcsT-110R: 5'-CCATTTCCATTTCACAGTTCG-3'. NHL1-880F and RbcsT-110R target the promoter and terminator sequences in the vector, so these could actually be used to genotype all of the seed. However, it must be noted that the latter primers cannot be used for RT-PCR analysis. The rbcSterm_FUE_R primer was designed because the annealing site is in the 5'-end of the terminator sequence that is included in the transcribed mRNA, hence it's application for RT-PCR.

Total RNA was isolated using the Nucleospin miRNA Isolation of Small and Large RNA kit (Clontech) from the *H. glycines*-infected root tissue for all positive plants/lines from the respective events for pre-miR396d, e and f (Table 1). RNA yield and purity was assessed with a NanoDrop, and integrity with agarose gel electrophoresis. Also, the cDNA of the *GmUBI3* reference gene was successfully amplified from all samples used for diagnosis of transgene expression. Unfortunately for pre-miR396f samples, multiple bands were always amplified using either pre-miR396f_F or pre-miR396f_loop_F in combination with rbcSterm_FUE_R, even though these same primers worked very well for genotyping, and it wasn't possible to determine whether the transcripts from the transgene were being amplified. However, for both pre-miR396d and e, we were able to get specific cDNA amplification of the transgenes (see Table 1), and thus, reassured us that the *PGmNHL1* nematode-inducible promoter was successful at

transcribing the pre-miR396 transgenes. Though again we were unable to prove expression for all transgenes in all transgenic events, these findings provided enough motive to continue to the phenotyping for changes in susceptibility to *H. glycines*.

For the first phenotyping experiment, we tested events 4, 10 and 32 for pre-miR396e, and used 25 seeds from each of plants A and B, thus totaling 50 total seeds per event. Seeds were sewn and inoculated with *H. glycines* as described above. Once the first set of trifoliolate leaves emerged (approximately one week after inoculation – two weeks after sewing) leaf disks were collected from each plant and transferred to 48-well PCR plates (wells with 200-uL volumes). DNA was released from the leaf disks using the QuickExtract Plant DNA Extraction Solution (Epicentre), and 1 or 3-uL of supernatant (extract) was used as template for genomic PCR with Quick-Load® Taq 2X Master Mix (NEB). All samples were first tested for amplification of *GmUBI3* reference, and then were tested for the transgene. Samples that amplified the pre-miR396e transgene (using pre-miR396e_F and rbcSterm_FUE_R primers) were concluded as transgenic. Samples that amplified *GmUBI3* reference but not the pre-miR396e transgene were concluded as nulls. Although we were confident that our classification of transgenic and null was accurate, there is a remote possibility that not all of the samples classified as nulls were actually nulls (false negatives). However, we accepted our results, and proceeded with the phenotyping experiments by comparing the transgenic plants with the nulls. In the end, our genotyping resulted in 12 total replicates (i.e., plants) for event 4, 13 replicates for event 10, and 18 replicates for event 32, with 18 total replicates of wild type or null.

Thirty days after inoculation, females were collected from each plant using standard laboratory techniques, and were stored in a 50-mL flat bottom falcon tubes at

~4˚C. The total number of females per plant was estimated by first resuspending the females/debris with a 1-mL pipette tip cut at the end, and obtaining an average number of females per 3-mL, then calculating the number of females per 50-mL. We also classified the females per plant into three body size categories, small, medium or large, using the same method. Finally, all the eggs from each tube of females/debris were processed using standard laboratory techniques, and the number of eggs per female was estimated per plant.

Surprisingly, none of the three transgenic pre-miR396e events resulted in reduced female numbers compared to wild type (Fig. 2). In fact, event 4 even resulted in significantly higher female numbers compared to wild type (Fig. 2), but the meaning of this result is unknown, and probably not biologically relevant. Although the variation about the means is quite large, the expectation was that this transgenic system would dramatically reduce the number of females, so the nearly identical means that we observed suggests that this system is not as effective at reducing female numbers as we'd planned.

We next evaluated the *H. glycines* females to test for differences in body size between transgenic pre-miR396e and wild type plants. Event 10 females were essentially identical with wild type (Fig. 3). We were at first very interested in finding that event 4 resulted in significantly smaller females compared to wild type, but our interest was immediately withdrawn when observing the opposite effect for event 32 (Fig. 3). Thus, overall, there really did not appear to be consistent differences in *H. glycines* female body sizes between the transgenic pre-miR396e events and wild type.

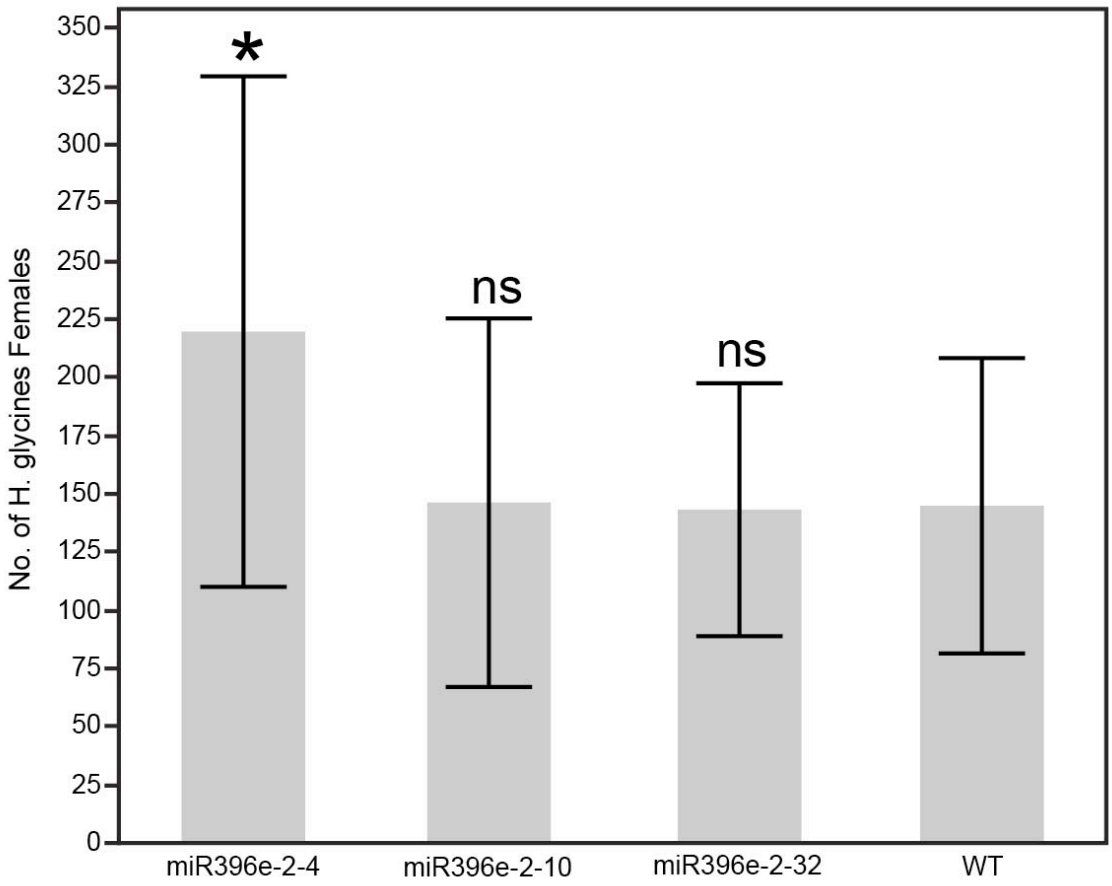



**Fig. 2. Number of *H. glycines* females that developed on the roots of transgenic pre-miR396e and wild type soybean plants.** Statistical comparisons between the transgenic events and wild type (WT) were performed with a student's t-test. Error bars represent one standard deviation from the mean. *, $P < 0.05$; ns, not significant.

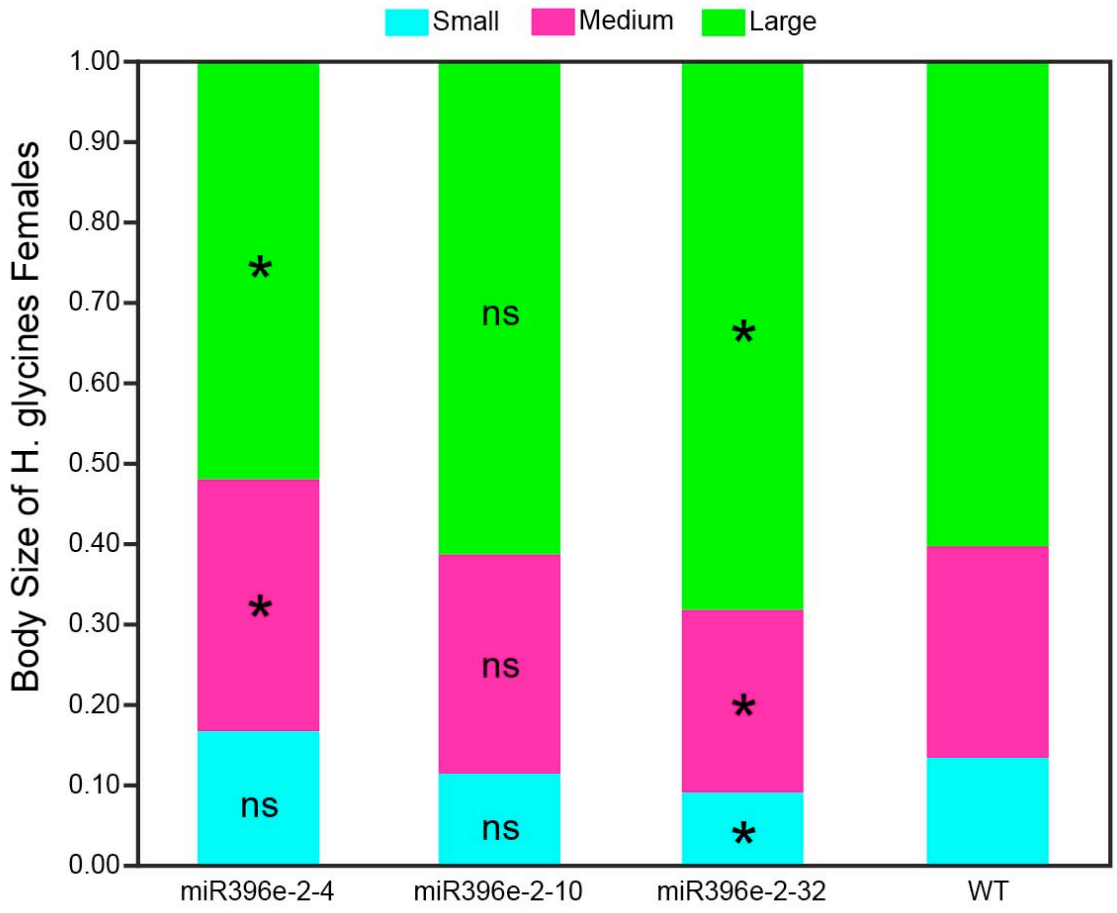



**Fig. 3. Body size comparisons of *H. glycines* females between transgenic pre-miR396e and wild type soybean plants.** Statistical comparisons were performed as in Figure 2.

Lastly, we tested whether the *H. glycines* females that developed on the transgenic pre-miR396e events produced different numbers of eggs compared to wild type, a measure of fecundity. This comparison was a bit more promising than the previous two comparisons as all three transgenic pre-miR396e events resulted in at least mild reductions in eggs per female compared to wild type (Fig. 4). However, only event 10 resulted in significantly reduced fecundity compared to wild type. Thus, although we saw mild reductions in fecundity, these results still were not promising.

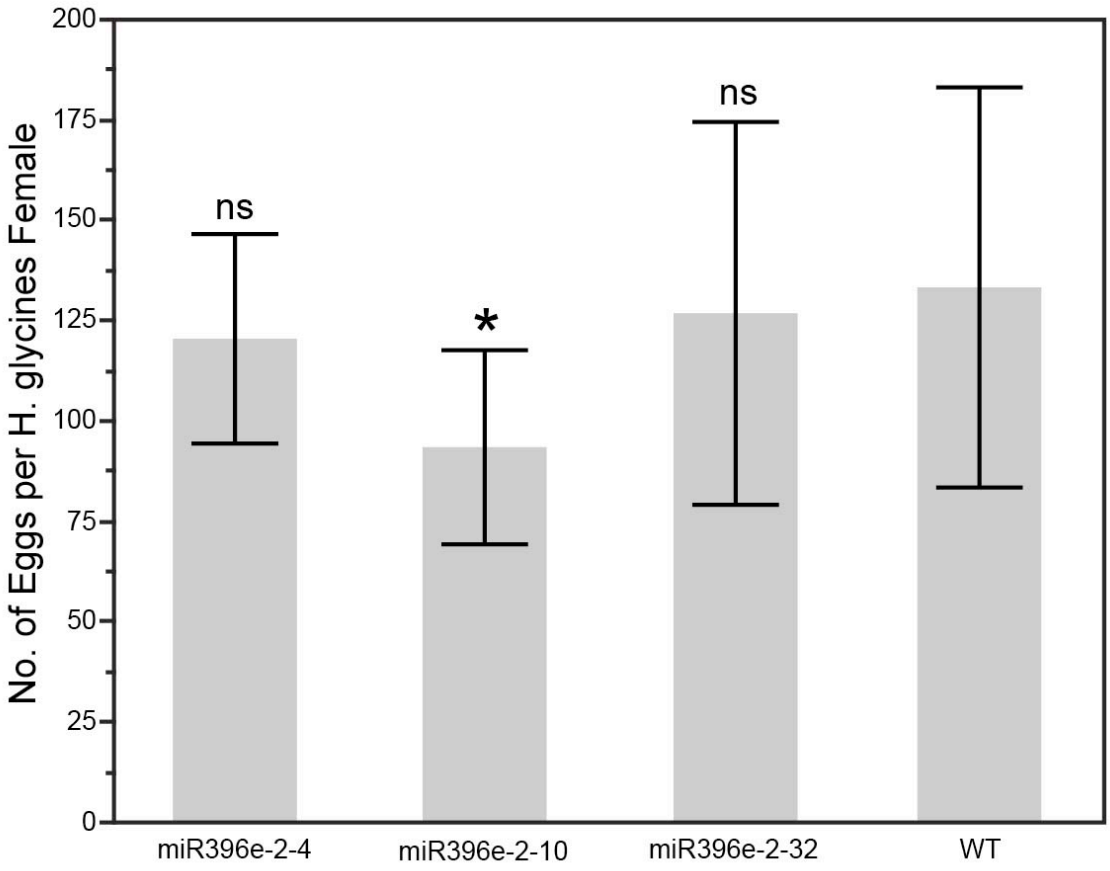



**Fig. 4. Number of eggs produced per *H. glycines* female that developed on pre-miR396e transgenic and wild type soybean plants.** Statistical comparisons were performed as in Figure 2.

The results obtained thus far for the transgenic pre-miR396e events are troubling for a number of reasons. Constitutive overexpression of all eight functional pre-miR396 in soybean hairy roots resulted in the same *EXTRA HAIRY* phenotype, reduced the induction of the *GRF* target genes during the syncytium formation phase, and consequently, resulted in substantially reduced numbers of *H. glycines* females compared to control. Thus, according to that data, all pre-miR396 are redundant (i.e., it shouldn't matter which functional pre-miR396 is used in our *PGmNHL1* transgenic system). This means that if the results shown above for pre-miR396e are reliable, which really all that

we have shown is that there isn't much difference from wild type, than the pre-miR396b, d and f transgenic events are unlikely to show any significant difference either. However, below I provide a number of possible, although unlikely, issues with either the pre-miR396e transgenic events, or with the experiments themselves. Then, with the assumption that there aren't technical issues with pre-miR396e, I provide a number of ideas for optimizing nematode-inducible expression of pre-miR396 to control for *H. glycines*.

Here are four possible technical issues with either the pre-miR396e transgenic events, or with the experimental design: Firstly, I have not PCR-amplified the *PGmNHL1:pre-miR396e:rbcSterm* construct from the transgenic events, cloned and sequenced in order to confirm that there are no sequence anomalies, so this remains as a possibility. To do this, I would use the NHL1-880F and RbcsT-110R primers for the genomic PCR. Secondly, although I was able to obtain evidence that likely indicates that the pre-miR396e transgene is expressed under nematode infection, I only obtained this evidence for one of the two plants for each event, but this is unlikely an issue. Thirdly, I used the T1 seed for the experiments, and it might be possible that homozygotes would show stronger phenotypes possibly through higher transgene expression, but I'm not very confident in this making much difference. Nonetheless, it may be worthwhile to obtain T3 homozygous seed for experimentation. Lastly, and probably most important, I was unhappy with the QuickExtract solution for isolating gDNA from leaf disks for genotyping. I found that this solution releases very low amounts of gDNA, and due to the large volume of solution added, the gDNA is very dilute for the PCR, and you can't simply add larger volumes because the extract contains plant compounds that are

inhibitory to the PCR reaction. So although I did see the presence or absence of bands from the genomic PCRs, the intensities of the bands were rather weak. This makes it possible that the samples concluded as nulls (nontransgenic), and thus were designated as wild types in the experiments, were actually false negatives, which would have completely ruined the reliability of the experiments above. However, I don't suspect that the plants concluded as nulls were false negatives—it is just a possibility. I would suggest using a different method of gDNA isolation from the leaf disks in future genotyping/phenotyping experiments. The ChargeSwitch gDNA Plant Kit worked very well for the preliminary genotyping analyses, and although this kit is more expensive and takes longer to use compared with QuickExtract, I believe it would provide more robust classifications of transgenic and null plants.

Assuming that the potential technical issues described above are not the problem, here are two possible biological issues that could explain the lack of robust phenotypes for the pre-miR396e transgenic events, and help to optimize the system: Firstly, *PGmNHL1* is only one of several different candidate nematode-inducible promoters identified by Kandoth et al. (2011). Since there haven't been any quantitative comparisons made between the activities of these nematode-inducible promoters, just simply promoter-reporter fusions for qualitative analyses, it is possible that *PGmNHL1* does not have strong enough activity within the syncytium to greatly impact syncytium development. In chapter 6, we showed that even with very strong constitutive overexpression of pre-miR396, *GRFs* still are induced at appreciable levels during syncytium formation. Thus, a quantitative comparison between all of the candidate nematode-inducible promoters might be essential in order to identify the best promoter to

use for this system. Secondly, in addition to the potential issues with the *PGmNHL1* promoter, we are using the natural forms of the pre-miR396 molecules. When bound to the *GRF* mRNA target, the natural miR396 contain a single bulge within the target site. It has been shown in Arabidopsis that when this bulge is removed, *GRF* silencing is much more robust (Debernardi et al. 2012). Thus, the most efficacious strategy may be to use an artificial pre-miR396 that will produce miR396 molecules that have perfect complementarity with *GRFs* for more robust silencing in syncytia, in combination with a nematode-inducible promoter with strongest, syncytium-specific activity.

## REFERENCES

Abad P et al. 2008. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nat Biotechnol. 26:909-915.

Ali S et al. 2015a. Analysis of putative apoplastic effectors from the nematode, *Globodera rostochiensis*, and identification of an expansin-like protein that can induce and suppress host defenses. PLoS One. 10:e0115042.

Ali S et al. 2015b. Analysis of *Globodera rostochiensis* effectors reveals conserved functions of SPRYSEC proteins in suppressing and eliciting plant immune responses. Front Plant Sci. 6:623.

Alkharouf NW et al. 2006. Timecourse microarray analyses reveal global changes in gene expression of susceptible *Glycine max* (soybean) roots during infection by *Heterodera glycines* (soybean cyst nematode). Planta. 224:838-852.

Altschul SF et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Antonino de Souza Junior JD et al. 2013. Knocking-down *Meloidogyne incognita* proteases by plant-delivered dsRNA has negative pleiotropic effect on nematode vigor. PLoS One. 8:e85364.

Ausubel FM. 2005. Are innate immune signaling pathways in plants and animals conserved? Nat Immunol. 6:973-979.

Baldwin JG, Nadler SA, Adams BJ. 2004. Evolution of plant parasitism among nematodes. Annu Rev Phytopathol. 42:83-105.

Balakirev MY et al. 2003. Otubains: A new family of cysteine proteases in the ubiquitin pathway. EMBO Rep. 4:517-522.

Banfield DK. 2011. Mechanisms of protein retention in the golgi. Cold Spring Harb Perspect Biol. 3:a005264.

Bao ML et al. 2014. miR396a-mediated basic helix-loop-helix transcription factor bHLH74 repression acts as a regulator for root growth in Arabidopsis seedlings. Plant Cell Physiol. 55:1343-1353.

Baum TJ, Wubben MJE, Hardy KA, Su H, Rodermel SR. 2000. A screen for *Arabidopsis thaliana* mutants with altered susceptibility to *Heterodera schachtii*. J Nematol. 32:166-173.

Bazin J et al. 2013. miR396 affects mycorrhization and root meristem activity in the legume *Medicago truncatula*. Plant J. 74:920-934.

Bekal S, Domier LL, Niblack TL, Lambert KN. 2011. Discovery and initial analysis of novel viral genomes in the soybean cyst nematode. J Gen Virol. 92:1870-1879.

Bekal S, Niblack TL, Lambert KN. 2003a. A chorismate mutase from the soybean cyst nematode *Heterodera glycines* shows polymorphisms that correlate with virulence. Mol Plant Microbe In. 16:439-446.

Bekal S, Niblack TL, Lambert KN. 2003b. Chorismate mutase (HG-CM) of the soybean cyst nematode, *Heterodera glycines*, causes altered root development when expressed in SCN resistant soybean hairy roots. J Nematol. 35:325-325.

Benkert P, Tosatto SCE, Schomburg D. 2008. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 71:261-277.

Bentley R. 1990. The shikimate pathway - a metabolic tree with many branches. Crit Rev Biochem Mol Biol. 25:307-384.

Bewley MC et al. 2006. Structures of wild-type and mutant human spermidine/spermine N-1-acetyltransferase, a potential therapeutic drug target. Proc Natl Acad Sci USA. 103:2063-2068.

Blaxter M, Koutsovoulos G. 2015. The evolution of parasitism in Nematoda. Parasitology. 142:S26-S39.

Blin N, Stafford DW. 1976. General method for isolation of high molecular-weight DNA from eukaryotes. Nucleic Acids Res. 3:2303-2308.

Bologna NG, Voinnet O. 2014. The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. Annu Rev Plant Biol. 65:473-503.

Boratyn GM et al. 2012. Domain enhanced lookup time accelerated BLAST. Biol Direct. 7:12.

Brown JWS, Shaw PJ. 2008. The role of the plant nucleolus in pre-mRNA processing. Curr Top Microbiol Immunol. 326:291-311.

Burnham KP, Anderson DR. 1998. *Model selection and inference: A practical information-theoretic approach*. Springer.

Cabrera J et al. 2016. Differentially expressed small RNAs in Arabidopsis galls formed by *Meloidogyne javanica*: a functional role for miR390 and its TAS3-derived tasiRNAs. New Phytol. 209:1625-1640.

Casadevall R, Rodriguez RE, Debernardi JM, Palatnik JF, Casati P. 2013. Repression of *Growth Regulating Factors* by the microRNA396 inhibits cell proliferation by UV-B radiation in Arabidopsis leaves. Plant Cell. 25:3570-3583.

Chakrabarty R et al. 2007. pSITE vectors for stable integration or transient expression of autofluorescent protein fusions in plants: Probing *Nicotiana benthamiana*-virus interactions. Mol Plant Microbe In. 20:740-750.

Chakravarthy S, Velasquez AC, Martin GB. 2009. Assay for pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI) in plants. J Vis Exp. 31:1442.

Chen CL et al. 2015. An ANNEXIN-like protein from the cereal cyst nematode *Heterodera avenae* suppresses plant defense. PLoS One. 10:e0122256.

Chen HM et al. 2010. 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. Proc Natl Acad Sci USA. 107:15269-15274.

Chen SY, Chronis D, Wang XH. 2013. The novel GrCEP12 peptide from the plant-parasitic nematode *Globodera rostochiensis* suppresses flg22-mediated PTI. Plant Signal Behav. 8:e25359.

Chen SY et al. 2001. Soybean cyst nematode population development and associated soybean yields of resistant and susceptible cultivars in Minnesota. Plant Dis. 85:760-766.

Chen VB et al. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 66:12-21.

Chook YM, Ke HM, Lipscomb WN. 1993. Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog. Proc Natl Acad Sci USA. 90:8600-8603.

Chronis D et al. 2013. A ubiquitin carboxyl extension protein secreted from a plant-parasitic nematode *Globodera rostochiensis* is cleaved in planta to promote plant parasitism. Plant J. 74:185-196.

Coll NS, Epple P, Dangl JL. 2011. Programmed cell death in the plant immune system. Cell Death Differ. 18:1247-1256.

Conley SP, Gaska JM, Pedersen P, Esker P. 2011. Soybean yield and *Heterodera Glycines* response to rotation, tillage, and genetic resistance. Agron J. 103:1604-1609.

Cook DE et al. 2014. Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. Plant Physiol. 165:630-647.

Cook DE et al. 2012. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science. 338:1206-1209.

Coppi A et al. 2011. The malaria circumsporozoite protein has two functional domains, each with distinct roles as sporozoites journey from mosquito to mammalian host. J Exp Med. 208:341-356.

Corpet F. 1988. Multiple sequence alignment with hierarchical-clustering. Nucleic Acids Res. 16:10881-10890.

Cotton JA et al. 2014. The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode. Genome Biol. 15:R43.

Craig, JP, Bekal S, Niblack T, Domier L, Lambert KN. 2009. Evidence for horizontally transferred genes involved in the biosynthesis of vitamin B1, B5, and B7 in *Heterodera glycines*. J Nematol. 41:281-290.

Craig JP et al. 2008. Analysis of a horizontally transferred pathway involved in vitamin B6 biosynthesis from the soybean cyst nematode *Heterodera glycines*. Mol Biol Evol. 25:2085-2098.

Dai X, Zhao PX. 2011. psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. 39:W155-W159.

Danchin EGJ et al. 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. Proc Natl Acad Sci USA. 107:17651-17656.

Davis EL, Hussey RS, Baum TJ. 2004. Getting to the roots of parasitism by nematodes. Trends Parasitol. 20:134-141.

Davis EL, Hussey RS, Baum TJ, Bakker J, Schots A. 2000. Nematode parasitism genes. Annu Rev Phytopathol. 38:365-396.

de Boer JM, Yan Y, Smant G, Davis EL, Baum TJ. 1998. In-situ hybridization to messenger RNA in *Heterodera glycines*. J Nematol. 30:309-312.

Debernardi JM et al. 2014. Post-transcriptional control of GRF transcription factors by microRNA miR396 and GIF co-activator affects leaf size and longevity. Plant J. 79:413-426.

Debernardi JM, Rodriguez RE, Mecchia MA, Palatnik JF. 2012. Functional specialization of the plant miR396 regulatory network through distinct microRNA-target interactions. PLoS Genet. 8:e1002419.

de la Chaux N, Wagner A. 2011. BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol. 11:154.

Dennis S, Sheth U, Feldman JL, English KA, Priess JR. 2012. *C. elegans* germ cells show temperature and age-dependent expression of *Cer1*, a *Gypsy/Ty3*-related retrotransposon. PLoS Pathog. 8:e1002591.

Di Cera E. 2009. Serine proteases. IUBMB Life. 61:510-515.

Ding X, Shields J, Allen R, Hussey RS. 1998. A secretory cellulose-binding protein cDNA cloned from the root-knot nematode (*Meloidogyne incognita*). Mol Plant Microbe In. 11:952-959.

Djamei A et al. 2011. Metabolic priming by a secreted fungal effector. Nature. 478:395-398.

Dong K, Barker KR, Opperman CH. 2005. Virulence genes in *Heterodera glycines*: Allele frequencies and ror gene groups among field isolates and inbred lines. Phytopathology. 95:186-191.

Doyle EA, Lambert KN. 2003. *Meloidogyne javanica* chorismate mutase 1 alters plant cell development. Mol Plant Microbe In. 16:123-131.

Draker KA, Wright GD. 2004. Molecular mechanism of the enterococcal aminoglycoside 6'-N-acetyltransferase': Role of GNAT-conserved residues in the chemistry of antibiotic inactivation. Biochemistry. 43:446-454.

Duan CG et al. 2012. Suppression of Arabidopsis ARGONAUTE1-mediated slicing, transgene-induced RNA silencing, and DNA methylation by distinct domains of the Cucumber mosaic virus 2b protein. Plant Cell. 24:259-274.

Eamens AL, Smith NA, Curtin SJ, Wang MB, Waterhouse PM. 2009. The *Arabidopsis thaliana* double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. RNA. 15:2219-2235.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics. 14:755-763.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.

Elling AA, Davis EL, Hussey RS, Baum TJ. 2007. Active uptake of cyst nematode parasitism proteins into the plant cell nucleus. Int J Parasitol. 37:1269-1279.

Elling AA et al. 2009. Sequence mining and transcript profiling to explore cyst nematode parasitism. BMC Genomics. 10:58.

Elsworth B, Wasmuth J, Blaxter M. 2011. NEMBASE4: The nematode transcriptome resource. Int J Parasitol. 41:881-894.

Engler JdA, Gheysen G. 2013. Nematode-induced endoreduplication in plant host cells: why and how? Mol Plant Microbe In. 26:17-24.

Espinosa A, Alfano JR. 2004. Disabling surveillance: bacterial type III secretion system effectors that suppress innate immunity. Cell Microbiol. 6:1027-1040.

Eves-van den Akker S et al. 2014. The transcriptome of *Nacobbus aberrans* reveals insights into the evolution of sedentary endoparasitism in plant-parasitic nematodes. Genome Biol Evol. 6:2181-2194.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186-194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175-185.

Fabro G et al. 2011. Multiple candidate effectors from the oomycete pathogen *Hyaloperonospora arabidopsidis* suppress host plant immunity. PLoS Pathog. 7:e1002348.

Fang XL et al. 2013. Identification and comparative analysis of cadmium tolerance-associated miRNAs and their targets in two soybean genotypes. PLoS One. 8:e81471.

Fontana MF et al. 2011. Secreted bacterial effectors that inhibit host protein synthesis are critical for induction of the innate immune response to virulent *Legionella pneumophila*. PLoS Pathog. 7:e1001289.

Frevert U et al. 1998. Malaria circumsporozoite protein inhibits protein synthesis in mammalian cells. EMBO J. 17:3816-3826.

Fuhrman LE, Goel AK, Smith J, Shianna KV, Aballay A. 2009. Nucleolar protein suppress *Caenorhabditis elegans* innate immunity by inhibiting p53/CEP-1. PLoS Genet. 5:e1000657.

Gao B, Allen R, Davis EL, Baum TJ, Hussey RS. 2004. Molecular characterisation and developmental expression of a cellulose-binding protein gene in the soybean cyst nematode *Heterodera glycines*. Int J Parasitol. 34:1377-1383.

Gao BL et al. 2001. Identification of putative parasitism genes expressed in the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*. Mol Plant Microbe In. 14:1247-1254.

Gao BL et al. 2003. The parasitome of the phytonematode *Heterodera glycines*. Mol Plant Microbe In. 16:720-726.

Ganko EW, Fielman KT, McDonald JF. 2001. Evolutionary history of *Cer* elements and their impact on the *C. elegans* genome. Genome Res. 11:2066-2074.

Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science. 320:1210-1213.

Goverse A, Smant G. 2014. The activation and suppression of plant innate immunity by parasitic nematodes. Annu Rev Phytopathol. 52:243-265.

Graewe S, Stanway RR, Rennenberg A, Heussler VT. 2012. Chronicle of a death foretold: *Plasmodium* liver stage parasites decide on the fate of the host cell. FEMS Microbiol Rev. 36:111-130.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. 36:D154-D158.

Guo N et al. 2011. Microarray profiling reveals microRNAs involving soybean resistance to *Phytophthora sojae*. Genome. 54:954-958.

Guo Y, Ni J, Denver R, Wang X, Clark SE. 2011. Mechanisms of molecular mimicry of plant CLE peptide ligands by the parasitic nematode *Globodera rostochiensis*. Plant Physiol. 157:476-484.

Gutierrez L et al. 2009. Phenotypic plasticity of adventitious rooting in Arabidopsis is controlled by complex regulation of AUXIN RESPONSE FACTOR transcripts and microRNA abundance. Plant Cell. 21:3119-3132.

Haegeman A, Jones JT, Danchin EGJ. 2011a. Horizontal gene transfer in nematodes: A catalyst for plant parasitism? Mol Plant Microbe In. 24:879-887.

Haegeman A, Joseph S, Gheysen G. 2011b. Analysis of the transcriptome of the root lesion nematode *Pratylenchus coffeae* generated by 454 sequencing technology. Mol Biochem Parasitol. 178:7-14.

Hamamouch N et al. 2012. The interaction of the novel 30C02 cyst nematode effector protein with a plant beta-1,3-endoglucanase may suppress host defence to promote parasitism. J Exp Bot. 63:3683-3695.

Han Q, Robinson H, Ding HZ, Christensen BM, Li JY. 2012. Evolution of insect arylalkylamine N-acetyltransferases: Structural evidence from the yellow fever mosquito, *Aedes aegypti*. Proc Natl Acad Sci USA. 109:11669-11674.

HaouazineTakvorian N et al. 1997. Characterization of two members of the *Arabidopsis thaliana* gene family, At beta fruct3 and At beta fruct4 coding for vacuolar invertases. Gene. 197:239-251.

Hernandez-Garcia CM, Martinelli AP, Bouchard RA, Finer JJ. 2009. A soybean (*Glycine max*) polyubiquitin promoter gives strong constitutive expression in transgenic soybean. Plant Cell Rep. 28:837-849.

Hewezi T. 2015. Cellular signaling pathways and posttranslational modifications mediated by nematode effector proteins. Plant Physiol. 169:1018-1026.

Hewezi T, Baum TJ. 2012. Complex feedback regulations govern the expression of miRNA396 and its *GRF* target genes. Plant Signal Behav. 7:749-751.

Hewezi T, Baum TJ. 2013. Manipulation of plant cells by cyst and root-knot nematode effectors. Mol Plant Microbe In. 26:9-16.

Hewezi T, Baum TJ. 2015. Gene silencing in nematode feeding sites. Adv Bot Res. 73:221-239.

Hewezi T et al. 2006. Antisense expression of a NBS-LRR sequence in sunflower (*Helianthus annuus* L.) and tobacco (*Nicotiana tabacum* L.): evidence for a dual role in plant development and fungal resistance. Transgenic Res. 15:165-180.

Hewezi T et al. 2008a. Cellulose binding protein from the parasitic nematode *Heterodera schachtii* interacts with Arabidopsis pectin methylesterase: cooperative cell wall modification during parasitism. Plant Cell. 20:3080-3093.

Hewezi T et al. 2010. Arabidopsis spermidine synthase is targeted by an effector protein of the cyst nematode *Heterodera schachtii*. Plant Physiol. 152:968-984.

Hewezi T et al. 2015. The cyst nematode effector protein 10A07 targets and recruits host posttranslational machinery to mediate its nuclear trafficking and to promote parasitism in Arabidopsis. Plant Cell. 27:891-907.

Hewezi T, Howe P, Maier TR, Baum TJ. 2008b. Arabidopsis small RNAs and their targets during cyst nematode parasitism. Mol Plant Microbe In. 21:1622-1634.

Hewezi T, Maier TR, Nettleton D, Baum TJ. 2012. The Arabidopsis microRNA396-*GRF1/GRF3* regulatory module acts as a developmental regulator in the reprogramming of root cells during cyst nematode infection. Plant Physiol. 159:321-335.

Ho SN, Hunt HD, Horton RM, Pullen JK, Pease LR. 1989. Site-directed mutagenesis by overlap extension using the polymerase chain-reaction. Gene. 77:51-59.

Hogenhout SA, Van der Hoorn RAL, Terauchi R, Kamoun S. 2009. Emerging concepts in effector biology of plant-associated organisms. Mol Plant Microbe In. 22:115-122.

Holterman M et al. 2006. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Mol Biol Evol. 23:1792-1800.

Horiguchi G et al. 2011. ANGUSTIFOLIA3 plays roles in adaxial/abaxial patterning and growth in leaf morphogenesis. Plant Cell Physiol. 52:112-124.

Huang GZ et al. 2005. Two chorismate mutase genes from the root-knot nematode *Meloidogyne incognita*. Mol Plant Pathol. 6:23-30.

Huang GZ et al. 2004. Use of solid-phase subtractive hybridization for the identification of parasitism gene candidates from the root-knot nematode *Meloidogyne incognita*. Mol Plant Pathol. 5:217-222.

Huang GZ et al. 2003. A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode *Meloidogyne incognita*. Mol Plant Microbe In. 16:376-381.

Huang J. 2013. Horizontal gene transfer in eukaryotes: The weak-link model. BioEssays. 35:868-875.

Hugel FU, Pradel G, Frevert U. 1996. Release of malaria circumsporozoite protein into the host cell cytoplasm and interaction with ribosomes. Mol Biochem Parasitol. 81:151-170.

Huot B, Yao J, Montgomery BL, He SY. 2014. Growth-defense tradeoffs in plants: a balancing act to optimize fitness. Mol Plant. 7:1267-1287.

Hussey RS. 1985. Staining nematodes in plant tissue. In: Zuckerman BM, Mai WF, Harrison MB, editors. *Plant nematology laboratory manual*. The University of Massachusetts Agricultural Experiment Station. p. 197-199.

Hussey RS. 1989. Disease-inducing secretions of plant-parasitic nematodes. Annu Rev Phytopathol. 27:123-141.

Hussey RS, Grundler FMW. 1998. Nematode parasitism of plants. In: Perry RN, Wright DJ, editors. *The physiology and biochemistry of free-living and plant-parasitic nematodes*. Cab International. p. 213-243.

Ithal N et al. 2007a. Parallel genome-wide expression profiling of host and pathogen during soybean cyst nematode infection of soybean. Mol Plant Microbe In. 20:293-305.

Ithal N et al. 2007b. Developmental transcript profiling of cyst nematode feeding cells in soybean roots. Mol Plant Microbe In. 20:510-525.

Jacobson S, Pillus L. 1999. Modifying chromatin and concepts of cancer. Curr Opin Genet Dev. 9:175-184.

Jones JDG, Dangl JL. 2006. The plant immune system. Nature. 444:323-329.

Jones JT et al. 2003. Characterization of a chorismate mutase from the potato cyst nematode *Globodera pallida*. Mol Plant Pathol. 4:43-50.

Jones P et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236-1240.

Jones JT et al. 2013. Top 10 plant-parasitic nematodes in molecular plant pathology. Mol Plant Pathol. 14:946-961.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell. 14:787-799.

Kall L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. Nucleic Acids Res. 35:W429-W432.

Kandoth PK, Mitchum MG. 2013. War of the worms: how plants fight underground attacks. Curr Opin Plant Biol. 16:457-463.

Kandoth PK et al. 2011. The soybean Rhg1 locus for resistance to the soybean cyst nematode *Heterodera glycines* regulates the expression of a large number of stress- and defense-related genes in degenerating feeding cells. Plant Physiol. 155:1960-1975.

Kaplan I, Sardanelli S, Rehill BJ, Denno RF. 2011. Toward a mechanistic understanding of competition in vascular-feeding herbivores: an empirical test of the sink competition hypothesis. Oecologia. 166:627-636.

Katiyar-Agarwal S et al. 2006. A pathogen-inducible endogenous siRNA in plant immunity. Proc Natl Acad Sci USA. 103:18002-18007.

Kaushal D, Naeve CW. 2004. Loading and preparing data for analysis in spotfire. Curr Protoc Bioinformatics. Unit 7.8.

Kim JS et al. 2012. Arabidopsis GROWTH-REGULATING FACTOR7 functions as a transcriptional repressor of abscisic acid- and osmotic stress-responsive genes, including DREB2A. Plant Cell. 24:3393-3405.

Kim JH, Lee BH. 2006. GROWTH-REGULATING FACTOR4 of *Arabidopsis thaliana* is required for development of leaves, cotyledons, and shoot apical meristem. J Plant Biol. 49:463-468.

Kim KH et al. 2012. *Mycobacterium tuberculosis* Eis protein initiates suppression of host immune responses by acetylation of DUSP16/MKP-7. Proc Natl Acad Sci USA. 109:7729-7734.

Kim SH et al. 2007. Cajal bodies and the nucleolus are required for plant virus systemic infection. EMBO J. 26:2169-2179.

Kim YH, Kim KS, Riggs RD. 1986. Morphological-characteristics of syncytia in susceptible hosts infected by the soybean cyst nematode. Phytopathology. 76:913-917.

Klink VP, Hosseini P, Matsye P, Alkharouf NW, Matthews BF. 2009. A gene expression analysis of syncytia laser microdissected from the roots of the *Glycine max* (soybean) genotype PI 548402 (Peking) undergoing a resistant reaction after infection by *Heterodera glycines* (soybean cyst nematode). Plant Mol Biol. 71:525-567.

Koenning SR, Wrather, JA. 2010. Suppression of soybean yield potential in the continental United States by plant diseases from 2006 to 2009. Plant Health Prog. doi:10.1094/PHP-2010-1122-01-RS.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 42:D68-D73.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol. 305:567-580.

Kulcheski FR et al. 2011. Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics. 12:307.

Kumar M et al. 2014. De novo transcriptome sequencing and analysis of the cereal cyst nematode, *Heterodera avenae*. PLoS One. 9:e96311.

Lambert KN, Allen KD, Sussex IM. 1999. Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic nematode *Meloidogyne javanica*. Mol Plant Microbe In. 12:328-336.

Lee AHY et al. 2012. A bacterial acetyltransferase destroys plant microtubule networks and blocks secretion. PLoS Pathog. 8:e1002523.

Lee AY, Karplus PA, Ganem B, Clardy J. 1995. Atomic-structure of the buried catalytic pocket of *Escherichia coli* chorismate mutase. J Am Chem Soc. 117:3627-3628.

Lee C et al. 2011. The novel cyst nematode effector protein 19C07 interacts with the Arabidopsis auxin influx transporter LAX3 to control feeding site development. Plant Physiol. 155:866-880.

Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12:1048-1059.

Li H, Deng Y, Wu TL, Subramanian S, Yu O. 2010a. Misexpression of miR482, miR1512, and miR1515 increases soybean nodulation. Plant Physiol. 153:1759-1770.

Li WB et al. 2015. Identification of microRNAs in response to different day lengths in soybean using high-throughput sequencing and qRT-PCR. PLoS One. 10:e0132621.

Li XY et al. 2012. Identification of soybean microRNAs involved in soybean cyst nematode infection by deep sequencing. PLoS One. 7:e39650.

Li Y et al. 2010b. Identification of microRNAs involved in pathogen-associated molecular pattern-triggered plant innate immunity. Plant Physiol. 152:2222-2231.

Liang G, He H, Li Y, Wang F, Yu D. 2014. Molecular mechanism of microRNA396 mediating pistil development in Arabidopsis. Plant Physiol. 164:249-258.

Lilley CJ, Kyndt T, Gheysen G. 2011. Nematode resistant GM crops in industrialised and developing countries. In: Jones J, Gheysen G, Fenoll C, editors. *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Springer. p. 517-541.

Lin B et al. 2016. A novel nematode effector suppresses plant immunity by activating host reactive oxygen species-scavenging system. New Phytol. 209:1159-1173.

Liu HH, Tian X, Li YJ, Wu CA, Zheng CC. 2008. Microarray-based analysis of stress-regulated microRNAs in *Arabidopsis thaliana*. RNA. 14:836-843.

Liu J et al. 2012a. The BnGRF2 gene (GRF2-like gene from *Brassica napus*) enhances seed oil production through regulating cell number and plant photosynthesis. J Exp Bot. 63:3727-3740.

Liu JY, Rice JH, Chen NN, Baum TJ, Hewezi T. 2014. Synchronization of developmental processes and defense signaling by Growth Regulating Transcription Factors. PLoS One. 9:e98477.

Liu SM et al. 2012. A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. Nature. 492:256-260.

Liu Y et al. 2013. Response of tobacco to the *Pseudomonas syringae* pv. tomato DC3000 is mainly dependent on salicylic acid signaling pathway. FEMS Microbiol Lett. 344:77-85.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods. 25:402-408.

Lixin R, Efhymiadis A, Henderson B, Jans DA. 2001. Novel properties of the nucleolar targeting signal of human angiogenin. Biochem Biophys Res Commun. 284:185-193.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494-6506.

Lozano-Torres JL et al. 2012. Dual disease resistance mediated by the immune receptor Cf-2 in tomato requires a common virulence target of a fungus and a nematode. Proc Natl Acad Sci USA. 109:10119-10124.

Lozano-Torres JL et al. 2014. Apoplastic venom allergen-like proteins of cyst nematodes modulate the activation of basal plant innate immunity by cell surface receptors. PLoS Pathog. 10:e1004569.

Maafi ZT, Subbotin SA, Moens M. 2003. Molecular identification of cyst-forming nematodes (Heteroderidae) from Iran and a phylogeny based on ITS-rDNA sequences. Nematology. 5:99-111.

Maekawa T, Kufer TA, Schulze-Lefert P. 2011. NLR functions in plant and animal immune systems: so far and yet so close. Nat Immunol. 12:818-826.

Maier TR, Hewezi T, Peng JQ, Baum TJ. 2013. Isolation of whole esophageal gland cells from plant-parasitic nematodes for transcriptome analyses and effector identification. Mol Plant Microbe In. 26:31-35.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 10:1307-1318.

Mallory AC, Elmayan T, Vaucheret H. 2008. MicroRNA maturation and action - the expanding roles of ARGONAUTEs. Curr Opin Plant Biol. 11:560-566.

Manosalva P et al. 2015. Conserved nematode signalling molecules elicit plant defenses and pathogen resistance. Nat Commun. 6:7795.

Marchler-Bauer A, Bryant SH. 2004. CD-search: protein domain annotations on the fly. Nucleic Acids Res. 32:W327-W331.

Marchler-Bauer A et al. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43:D222-D226.

Marin E et al. 2010. miR390, Arabidopsis TAS3 tasiRNAs, and their AUXIN RESPONSE FACTOR targets define an autoregulatory network quantitatively regulating lateral root growth. Plant Cell. 22:1104-1117.

Matas IM et al. 2014. Translocation and functional analysis of *Pseudomonas savastanoi* pv. *savastanoi* NCPPB 3335 type III secretion system effectors reveals two novel effector families of the *Pseudomonas syringae* complex. Mol Plant-Microbe In. 27:424-436.

Mitchell A et al. 2015. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43:D213-D221.

Mitchum MG et al. 2013. Nematode effector proteins: an emerging paradigm of parasitism. New Phytol. 199:879-894.

Mitsui H et al. 2010. Phylogeny of Asian primate malaria parasites inferred from apicoplast genome-encoded genes with special emphasis on the positions of *Plasmodium vivax* and *P. fragile*. Gene. 450:32-38.

Nakai K, Horton P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci. 24:34-36.

Neveu C, Abad P, Castagnone-Sereno P. 2003. Molecular cloning and characterization of an intestinal cathepsin L protease from the plant-parasitic nematode *Meloidogyne incognita*. Physiol Mol Plant Pathol. 63:159-165.

Newman MA, Sundelin T, Nielsen JT, Erbs G. 2013. MAMP (microbe-associated molecular pattern) triggered immunity in plants. Front Plant Sci. 4:139.

Niblack TL, Heinz RD, Smith GS, Donald PA. 1994. Distribution, density and diversity of *Heterodera glycines* in Missouri. J Nematol. 25:880-886.

Nikolaidis N, Doran N, Cosgrove DJ. 2014. Plant expansins in bacteria and fungi: evolution by horizontal gene transfer and independent domain fusion. Mol Biol Evol. 2:376-386.

Noon JB et al. 2015. Eighteen new candidate effectors of the phytonematode *Heterodera glycines* produced specifically in the secretory esophageal gland cells during parasitism. Phytopathology. 105:1362-1372.

Oka Y et al. 2000. New strategies for the control of plant-parasitic nematodes. Pest Manag Sci. 56:983-988.

Omidbakhshfard MA, Proost S, Fujikura U, Mueller-Roeber B. 2015. Growth-Regulating Factors (GRFs): a small transcription factor family with important functions in plant biology. Mol Plant. 8:998-1010.

Pacheco MA et al. 2012. The origin of malarial parasites in orangutans. PLoS One. 7:e34990.

Pajoro A et al. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol. 15:R41.

Paquette N et al. 2012. Serine/threonine acetylation of TGF beta-activated kinase (TAK1) by *Yersinia pestis* YopJ inhibits innate immune signaling. Proc Natl Acad Sci USA. 109:12710-12715.

Patel N et al. 2010. A nematode effector protein similar to annexins in host plants. J Exp Bot. 61:235-248.

Pellegrini M, Marcotte EM, Yeates TO. 1999. A fast algorithm for genome-wide analysis of proteins with repeated sequences. Proteins. 35:440-446.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785-786.

Postma WJ et al. 2012. The effector SPRYSEC-19 of *Globodera rostochiensis* suppresses CC-NB-LRR-mediated disease resistance in plants. Plant Physiol. 160:944-954.

Prasad CS, Gupta S, Kumar H, Tiwari M. 2013. Evolutionary and functional analysis of fructose bisphosphate aldolase of plant parasitic nematodes. Bioinformation. 9:1-8.

Puigbo P, Bravo IG, Garcia-Vallve S. 2008. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). BMC Bioinformatics. 9:65.

Qin L et al. 2000. An efficient cDNA-AFLP-based strategy for the identification of putative pathogenicity factors from the potato cyst nematode *Globodera rostochiensis*. Mol Plant Microbe In. 13:830-836.

Quevillon E et al. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33:W116-W120.

Quist CW, Smant G, Helder J. 2015. Evolution of plant parasitism in the phylum Nematoda. Annu Rev Phytopathol. 53:289-310.

Replogle A et al. 2011. Nematode CLE signaling in Arabidopsis requires CLAVATA2 and CORYNE. Plant J. 65:430-440.

Rodriguez RE et al. 2015. MicroRNA miR396 regulates the switch between stem cells and transit-amplifying cells in Arabidopsis roots. Plant Cell. 27:3354-3366.

Rodriguez RE et al. 2010. Control of cell proliferation in *Arabidopsis thaliana* by microRNA miR396. Development. 137:103-112.

Rogers K, Chen XM. 2013. Biogenesis, turnover, and mode of action of plant microRNAs. Plant Cell. 25:2383-2399.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 5:725-738.

Rutter WB et al. 2014. Mining novel effector proteins from the esophageal gland cells of *Meloidogyne incognita*. Mol Plant Microbe In. 27:965-974.

Sanches M, Alves BSC, Zanchin NIT, Guimaraes BG. 2007. The crystal structure of the human Mov34 MPN domain reveals a metal-free dimer. J Mol Biol. 370:846-855.

Schmutz J et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature. 463:178-183.

Schoenknecht G et al. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science. 339:1207-1210.

Scholl EH, Thorne JL, McCarter JP, Bird DM. 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol. 4:R39.

Shamimuzzaman M, Vodkin L. 2012. Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. BMC Genomics. 13:310.

Shaw P, Brown J. 2012. Nucleoli: composition, function, and dynamics. Plant Physiol. 158:44-51.

Sindhu AS et al. 2009. Effective and specific in planta RNAi in cyst nematodes: expression interference of four parasitism genes reduces parasitic success. J Exp Bot. 60:315-324.

Singh AP et al. 2007. *Plasmodium* circumsporozoite protein promotes the development of the liver stages of the parasite. Cell. 131:492-504.

Smant G et al. 1998. Endogenous cellulases in animals: Isolation of beta-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc Natl Acad Sci USA. 95:4906-4911.

Sobezak M, Golinowski W. 2009. Structure of cyst nematode feeding sites. In: Berg RH, Taylor CG, editors. *Plant Cell Monographs*. Springer. p. 153-187.

Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 6:175-182.

Spoel SH, Dong XN. 2012. How to plants achieve immunity? Defense without specialized immune cells. Nat Rev Immunol. 12:89-100.

Subbotin S, Mundo-Ocampo M, Baldwin JG. 2010. Systematics of cyst nematodes (Nematoda: Heteroderinae). In: Hunt DJ, Perry RN, editors. *Nematology Monographs and Perspectives*. p. 1-512.

Sun BF et al. 2013. Multiple ancient horizontal gene transfers and duplications in lepidopteran species. Insect Mol Biol. 1:72-87.

Sun XL, Rikkerink EHA, Jones WT, Uversky VN. 2013. Multifarious roles of intrinsic disorder in proteins illustrate its broad impact on plant biology. Plant Cell. 25:38-55.

Szakasits D et al. 2009. The transcriptome of syncytia induced by the cyst nematode *Heterodera schachtii* in Arabidopsis roots. Plant J. 57:771-784.

Szklarczyk R, Heringa J. 2004. Tracking repeats using significance and transitivity. Bioinformatics. 20:311-317.

Tamura K et al. 2011. MEGA5: molecular evolutionary genetics analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony methods. Mol Biol Evol. 28:2731-2739.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 30:2725-2729.

Theobald DL. 2010. A formal test of the theory of universal common ancestry. Nature. 465:219-222.

Thorpe P et al. 2014. Genomic characterisation of the effector complement of the potato cyst nematode *Globodera pallida*. BMC Genomics. 15:923.

Toubarro D et al. 2010. Serine protease-mediated host invasion by the parasitic nematode *Steinernema carpocapsae*. J Biol Chem. 285:30666-30675.

Tytgat T et al. 2004. A new class of ubiquitin extension proteins secreted by the dorsal pharyngeal gland in plant parasitic cyst nematodes. Mol Plant Microbe In. 17:846-852.

van der Knaap E, Kim JH, Kende H. 2000. A novel gibberellin-induced gene from rice and its potential regulatory role in stem growth. Plant Physiol. 122:695-704.

Van Z et al. 2013. miR172 regulates soybean nodulation. Mol Plant Microbe In. 26:1371-1377.

Vanholme B et al. 2007. Molecular characterization and functional importance of pectate lyase secreted by the cyst nematode *Heterodera schachtii*. Mol Plant Pathol. 8:267-278.

Varmus H, Brown P. 1989. Retroviruses. In: Berg DE, Howe MM, editors. *Mobile DNA*. American Society for Microbiology. p. 53-108.

Veronico P, Jones J, Di Vito M, De Giorgi C. 2001. Horizontal transfer of a bacterial gene involved in polyglutamate biosynthesis to the plant-parasitic nematode *Meloidogyne artiellia*. FEBS Lett. 508:470-474.

Vetting MW et al. 2005. Structure and functions of the GNAT superfamily of acetyltransferases. Arch Biochem Biophys. 433:212-226.

Vitha S, Benes K, Phillips JP, Gartland KMA. 1995. Histochemical GUS analysis. In: Gartland KMA, Davey MR, editors. *Methods in Molecular Biology; Agrobacterium protocols*. Humana Press Inc. p. 185-193.

Wagstaff KM, Jans DA. 2009. Importins and beyond: non-conventional nuclear transport mechanisms. Traffic. 10:1188-1198.

Wang J et al. 2010. Dual roles for the variable domain in protein trafficking and host-specific recognition of *Heterodera glycines* CLE effector proteins. New Phytol. 187:1003-1017.

Wang J et al. 2011. Identification of potential host plant mimics of CLAVATA3/ESR (CLE)-like peptides from the plant-parasitic nematode *Heterodera schachtii*. Mol Plant Pathol. 12:177-186.

Wang JW, Czech B, Weigel D. 2009. miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. Cell. 138:738-749.

Wang L, Brown SJ. 2006. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 34:W243-W248.

Wang R, Brattain MG. 2007. The maximum size of protein to diffuse through the nuclear pore is larger than 60 kDa. FEBS Lett. 581:3164-3170.

Wang XH et al. 2005. A parasitism gene from a plant-parasitic nematode with function similar to CLAVATA3/ESR (CLE) of *Arabidopsis thaliana*. Mol Plant Pathol. 6:187-191.

Wang XH et al. 2001. Signal peptide-selection of cDNA cloned directly from the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*. Mol Plant Microbe In. 14:536-544.

Wang XM et al. 2013. Characterization of a novel NBS-LRR gene involved in bacterial blight resistance in rice. Plant Mol Biol Rep. 31:649-656.

Wang YJ et al. 2013. Elucidation of miRNAs-mediated responses to low nitrogen stress by deep sequencing of two soybean genotypes. PLoS One. 8:e67423.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25:1189-1191.

Weiberg A et al. 2013. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. Science. 342:118-123.

Wheeler DL et al. 2008. Database resources of the national center for biotechnology information. Nucleic Acids Res. 36:D13-D21.

Wijayawardena BK, Minchella DJ, DeWoody JA. 2013. Hosts, parasites, and horizontal gene transfer. Trends Parasitol. 29:329-338.

Wirthmueller L, Roth C, Banfield MJ, Wiermer M. 2013. Hop-on hop-off: importin-alpha guided tours to the nucleus in innate immune signaling. Front Plant Sci. 4:149.

Wong CE et al. 2011. MicroRNAs in the shoot apical meristem of soybean. J Exp Bot. 62:2495-2506.

Wright S, Finnegan D. 2001. Genome evolution: sex and the transposable element. Curr Biol. 11:R296-R299.

Wu G et al. 2009. The Sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. Cell. 138:750-759.

Wu HX, Jones RM, Neish AS. 2012. The Salmonella effector AvrA mediates bacterial intracellular survival during infection in vivo. Cell Microbiol. 14:28-39.

Wylie T et al. 2004. Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes. Nucleic Acids Res. 32:D423-D426.

Xu F et al. 2013. Genome-wide identification of soybean microRNAs and their targets reveals their organ-specificity and responses to phosphate starvation. BMC Genomics. 14:66.

Xu MY et al. 2014. Novel miRNA and phasiRNA biogenesis networks in soybean roots from two sister lines that are resistant and susceptible to SCN race 4. PLoS One. 9:e110051.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35:W265-W268.

Yan Z et al. 2015. Identification of microRNAs and their mRNA targets during soybean nodule development: functional analysis of the role of miR393j-3p in soybean nodulation. New Phytol. 207:748-759.

Yant L et al. 2010. Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2. Plant Cell. 22:2156-2170.

Yu B et al. 2005. Methylation as a crucial step in plant microRNA biogenesis. Science. 307:932-935.

Yu H, Chronis D, Lu SW, Wang XH. 2011. Chorismate mutase: an alternatively spliced parasitism gene and a diagnostic marker for three important *Globodera* nematode species. Eur J Plant Pathol. 129:89-102.

Yu J et al. 2014. SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways. BMC Genomics. 15:271.

Zhang X et al. 2011. Arabidopsis Argonaute 2 regulates innate immunity via miRNA393*-mediated silencing of a Golgi-localized SNARE gene, MEMB12. Mol Cell. 42:356-366.

Zhao C, Doucet D, Mittapalli O. 2014. Characterization of horizontally transferred β-fructofuranosidase (ScrB) genes in *Agrilus planipennis*. Insect Mol Biol. 6:821-832.

Zhao MX, Meyers BC, Cai CM, Xu W, Ma JX. 2015a. Evolutionary patterns and coevolutionary consequences of *MIRNA* genes and microRNA targets triggered by multiple mechanisms of genomic duplications in soybean. Plant Cell. 27:546-562.

Zhao W et al. 2015b. Identification of jasmonic acid-associated microRNAs and characterization of the regulatory roles of the miR319/TCP4 module under root-knot nematode stress in tomato. J Exp Bot. 66:4653-4667.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406-3415.